AGENCY, MANAGED CARE AND FINANCIAL-RISK SHARING IN GENERAL MEDICAL PRACTICE

Cover: fragment of painting by Eelkjen Kloos
Layout: Titia Kloos
Printed by: Gildeprint B.V., Enschede

AGENCY, MANAGED CARE AND FINANCIAL-RISK SHARING IN GENERAL MEDICAL PRACTICE

Agentschapsrelaties, 'managed care' en financiële-risicodeling
in de huisartsenzorg

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus prof.dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 14 september 2006 om 13.30 uur

door

Ad Vermaas
geboren te Utrecht

Promotiecommissie

Promotoren:	Prof.dr. F.T. Schut Prof.dr. W.P.M.M. van de Ven
Overige leden:	Prof.dr. G.W.J. Hendrikse Prof.dr. H.A. Keuzenkamp Prof.dr. W.A.B. Stalman

CONTENTS

1 INTRODUCTION

1.1 Introduction to the research

In several health-care systems general practitioners have a prominent position. Although their position, role and functioning may differ per system, general practitioners (GPs) or primary care physicians (like family practitioners, general paediatricians and non-subspecialising internists) provide direct accessible, comprehensive, longitudinal or continuous and personal care (Starfield 1992, Fry and Horder 1994). As distinguished from medical specialists, GPs often form the first point of contact for patients with the health-care system. This is especially the case if they have a role as gatekeeper (Boerma et al. 1997). The gatekeeper function involves that patients only have access to other providers of care after they have obtained a referral from their GP. It is the GP then who ultimately decides whether a patient can visit a medical specialist or whether he can visit one without being confronted with cost sharing. Further, GPs often have a role as the patients' guide. As a guide, the GP ensures that the patient is referred to the right provider of care at the right time.

An important rationale for their role as guide or gatekeeper is that GPs are supposed to have superior information about health and diseases, diagnostics, treatments, the health-care system and the quality of other providers of care. This provides them with an interesting position. GPs are providers of health-care services and are positioned at the supply-side of the health-care market. Although patients are at the demand-side of this market, they initiate only part of the demand for health services by themselves. Repeated contacts and the contacts with other providers of care will often be initiated by the GP. GPs may therefore be viewed as suppliers as well as demanders of care (Schut 1999). As a guide or gatekeeper, the GP is supposed to use his superior information to induce the demand for care. This demand-inducement should protect the ill-informed patient from demanding too much or too little goods and services. Hence a GP that influences the patients' decisions about the use of health care, or even is the decision-maker himself, has a large influence on the nature, the quantity, the quality and the costs of care.

While influencing the patient's demand for care, the GP is supposed to act in the patient's best interests, i.e. to act as an agent for the patient. The (sometimes large) variations among physicians or practices in the use of health care (like differences in referral rates) suggest at least some inefficiency and probably even differences in health outcomes though. There is evidence that these variations can only partly be explained by patient characteristics, like age, sex, health status, and socio-economic status. In an extensive research among GPs in the Netherlands, large differences were found between practices as well as between individual GPs in the number of contacts per patient per year. For instance, after correcting for patient characteristics, the 95% confidence intervals for practices ranged from 3.9 to 9.1 contacts per patient per year. After correcting for practice characteristics and type of computer software, the 95% confidence intervals for

practices still ranged from 5.1 to 8.3 contacts per patient per year (NIVEL/RIVM 2004). The differences between individual GPs were smaller than the differences between practices. Obviously, there are similarities within practices, like the same practice assistant. Here the 95% confidence intervals ranged from 5.4 to 7.6 contacts per patient per year and, after correcting for practice characteristics and type of computer software, from 5.6 to 7.8 contacts per patient per year.

Tamblyn et al. (2003) found that both physician characteristics and practice characteristics appear to be associated with the likelihood of prescribing new drugs as well as the utilization rates of new drugs. Relevant characteristics were, for instance, physician sex, specialty, medical school, years since graduation and practice location.

Eisenberg (2002) identified several factors that seem to influence medical decision making. These factors can be described in three general categories:

1. Factors based on the physician's own benefit, like the desire for income, the desire for a style of practice, physician characteristics, practice setting and the role of clinical leadership (educationally influential doctors).
2. Factors based on the patient's benefit, like the patient's economic well-being, clinical factors, patient demand, defensive medicine, patient characteristics and patient convenience.
3. Factors based on the benefit of society at large, like the constraints of society's limited resources.

Information plays also a crucial role in the GP's functioning as agent for the patient. Firstly, the GP may have insufficient information (for instance on the efficacy, the effectiveness or the efficiency of a specific treatment), or he may lack experience. Secondly, the medical profession as a whole may lack knowledge of a specific disease or may not know whether certain health-care activities are appropriate (Pauly 1978). Besides an absolute absence of particular scientific information, there may be an inferior diffusion of such information as well (Phelps 1992). The latter may result in the insufficiently informed GP, as mentioned above. Thirdly, the GP may purposely exploit his informational advantage in order to affect his own welfare. His welfare may be a function of, among other factors, his income and his workload (Scott 1997). Clearly, if there is a 'broad zone of uncertainty' (Evans 1984), the GP has some room for discretionary behaviour and may pursue his own interests. Especially if the GP who prescribed the care provides it as well, then there is some room for discretionary behaviour. The demand for GP services is not fully independent of its supply then. Yet also for care prescribed by the GP but provided by another health-care provider, the GP may attempt to induce demand (Van Doorslaer and Schut 1999).

The variances in practice style indicate that GPs do not always act as perfect agents for their patients, deliberately or not.[1] This suggests some room for improvement, which may pertain to the quality as well as to the costs of care. Attempts to improve GP care may come from the profession itself as well as from the patients. Providers of care may for instance promote scientific research, improve the diffusion of scientific information, de-

[1] A uniform practice style is no guarantee that GPs act in their patients' best interests, but at least it suggests a consensus amongst physicians about best practices.

velop guidelines and certify practices or specific skills. Patients may also contribute, for instance by stating their preferences, organising themselves into patient groups and raising funds for scientific research. But as the patient is uncertain about the moment, nature and amount of future health-care expenses, he may want to seek insurance to cover his financial risk. The demand for health insurance that may result from this uncertainty (and that, among other things, depends on the probability of the financial loss as well as on the size of the loss) is an important rationale for the presence of a so-called third party in health care. As health insurance lowers the costs of care or even reduces these to zero at the moment of consumption, the patient may not be interested in reducing costs or improving the cost-effectiveness of care. Under certain conditions, it will be the third party that has an interest in the costs or even in the quality of care, and that may attempt to influence the provision of health care. The question then is why and how the third party may want to influence the way care is provided.

In several countries a key element in the reform of the health-care delivery and financing systems is the design of the incentive system. Now that the restrictive government policy of cost containment proved to be unsuccessful in guaranteeing access to a basic package of health care, the use of incentive systems is seen as an effective way to increase the efficiency of the health-care sector (Sørensen and Grytten 2003). A prominent role in such incentive-based systems is often given to third parties, like public or private health insurers. In several ways third parties are stimulated to organise an efficient health-care system. One way is to make them financially responsible for a specified package of health-care goods and services for a defined group of members (i.e. insured) and for a certain period of time. In the Netherlands, for instance, a new health insurance act is currently being implemented. This act rearranges the organisation and financing of the health-care sector. Mutually competing third parties (i.e. health insurers) are provided with the incentives to act as agents on behalf of the insured by arranging affordable, efficient care of high quality. A key element of their role is that they have to act as negotiators with providers of care for the price, contents and organisation of the care (Tweede Kamer 2003-2004).

Because, as mentioned above, GPs have a large influence on the nature, the quantity, the quality and the costs of care, the third party may want to enter into contractual relationships with GPs in order to induce them to alter their behaviour. Hence besides exerting influence on GP care itself, third parties may for instance try to alter the GP's behaviour regarding diagnostics or drug prescriptions (follow-up care). Especially the GP's referral behaviour may be of interest because of the high follow-up costs, for instance associated with hospital care. A distinction can be made between two types of referrals though. So-called 'supplementary referrals' are a supplement to the GP's services, whereas 'alternative referrals' are referrals for care that the GP could have provided as well (Lurås 2004). The third party may primarily focus on the alternative referrals when trying to enhance the efficiency of the health-care sector. However, by providing the GP with additional means (like personnel) or by financing additional training, the third party may be able to reduce the number of supplementary referrals as well.

Although one can think of a set of possible techniques a third party may use to influence the behaviour of GPs, the use of financial incentives is certainly one of the most intriguing techniques. It is a way to stimulate the GP to choose the preferred actions by making these actions more attractive, i.e. changing the relative attractiveness of alternative actions. It is also a technique that is still, and will always remain under discussion. On the one hand, opponents may argue that physicians act as agents for their patients and will not be susceptible to financial incentives. Other opponents may argue that physicians (just like other people) are indeed susceptible to these incentives, that the resulting behaviour is potentially dangerous for patients and that such incentives should therefore be omitted from contracts. On the other hand, advocates may argue that financial incentives can, if used sensibly, stimulate certain physician behaviour whilst maintaining the professional or the physician's individual autonomy.

Stating that physicians will solely act in their patients' best interests can be considered unrealistic, as the physicians' utility functions will contain factors like income and leisure besides a factor medical ethics (Flierman 1991). Stating that incentives may induce potentially dangerous behaviour and that they should not be applied, is rather naive since there is no payment system without financial incentives. Moreover, whether the incentive as such is potentially dangerous, depends on the kind of behaviour that is being provoked. Incentives may be used to increase the provision of specific care, like preventive care, which may be in the interests of the patient. But the patient may also benefit from cost-reducing incentives. Franks et al. (1992), for instance, reviewed a number of studies that showed the inherent dangers of over-consumption. Finally, financial incentives may result in a more efficient health-care system. For the individual patient this might possibly reduce the quality of care in the short run, but in the long run it will contribute to the financial accessibility of care for that individual as well as for society. Yet financial incentives will have to be applied with care. Financial incentives may, for instance, result in a stronger cost awareness of physicians. In a research among Swedish physicians, Forsberg et al. (2001) found that a strong cost awareness was a negative predictor of quality of care (when rated by physicians). They argued that it is a difficult balancing act between cost considerations and the quality of care.

The use of financial incentives only makes sense if there is an effect of the incentive system on the behaviour of the physician (and, finally, on the outcome of the physician's actions). There are, however, only a small number of studies that analysed this effect. Moreover, these studies often suffer from methodological problems (Scott and Hall 1995). Financial incentives seem to be a key element in the apparent success of so-called managed care organisations, but patient selection effects, physician selection effects and missing variables may bias this effect (Hellinger 1996). Nevertheless, there are studies that, despite their limitations, clearly indicate that financial incentives affect physicians' behaviour (see for instance Hickson et al. 1987, Hemenway et al. 1990, Flierman 1991, Delnoy et al. 1992 and Sørensen and Grytten 2003).

Making the third party financially responsible is seen as a way to further an efficiently organised health-care system. The third party on its turn may use a comparable incentive system in order to stimulate desired behaviour by the contracted GPs. The GPs can be made financially responsible for the provision of care by shifting (part of) the third

party's risk to them. The risk may pertain to GP care only, but because of their specific role and position in the health-care system the financial responsibility may be extended to, for instance, the GP's prescription and referral behaviour. Combining the GP's gatekeeping function with financial accountability for a broader spectrum of treatments may also enhance integration of health care in case of fragmented health care financing systems (Jegers et al. 2002). Clearly, there is a resemblance between the incentives for the third party and the incentives for the GP, if also the latter is made financially responsible for a specified package of health-care goods and services for a defined group of members (i.e. patients) for a certain period of time.

1.2 Aims and contents

The contractual relationships between third parties and GPs may vary by the financial and organisational arrangements. One of the provisions may be that GPs are made financially responsible. The third party may want to shift the financial risk (partly) to a GP, which will presumably be the case if it pertains to care that is to a high degree under the GP's control. If the financial risk is only partly shifted to the GP, then the risk is shared.

Financial-risk sharing is one of the several incentives the third party may use to influence the GPs' behaviour. As noted, there are not many reliable studies that analysed the effect of incentive systems on the behaviour of physicians (and, finally, on the outcome of the physicians' actions). Also, little theoretical, conceptual research has been done. This holds especially for payment systems for GPs and for systems of financial-risk sharing. This is remarkable as the behaviour of physicians has an effect on the efficiency of the health-care system and as the efficiency of health-care systems is a worldwide issue.

The main purpose of our study is to construct a conceptual framework for systems of financial-risk sharing between third parties and GPs, to use this for the review of several examples of financial-risk sharing, and to discuss how systems of financial-risk sharing should be structured.

The central questions of this thesis are then:

Is there a rationale for financial-risk sharing between third-party agents and general practitioners?
And if so, how should systems of financial-risk sharing be structured?

By finding an answer to these two central questions, we aim at contributing to the literature on payment systems for GPs. However, financial incentives are one possible technique out of a set of potential techniques the third party may use to influence the GPs' behaviour. In health care (the use of) such a set of techniques is often designated as 'managed care'. We will return to this in chapter 5, but for the moment it is sufficient to note that managed care is a rather diffuse concept that is not clearly defined and that lacks a sound theoretical, conceptual framework. In an extensive review of the managed-care literature, published in the Handbook of Health Economics, Glied noted that there is no

single broadly accepted definition of the term managed care. Further, Glied argued that because there is a tremendous variation in the nature of the managed-care plans, it is difficult to assess theoretically as well as empirically the economics of managed care. Hence it may make more sense to think of managed care as a combination of several mechanisms, although these mechanisms have changed over time. According to Glied, economic theory and empirical research have not kept pace with the development of managed care. 'Research is needed to identify which characteristics of managed care generate economically meaningful differences in outcomes and which are only superficial' (Glied 2000, p. 745). As financial-risk sharing is one of the managed-care techniques, we will first try to provide managed care as a whole with a theoretical basis. In chapter 3 we will discuss agency theory, which we will use to develop a theoretical background for managed care. Hence we also aim at contributing to the literature on managed care in general, not only by identifying the different managed-care techniques (or mechanisms) but also by bringing these techniques in connection with each other.

The two central questions of this thesis have been split up into ten research questions to be dealt with in the pertinent chapters. In chapter 2 we will introduce the two central parties in this thesis: the third party and the GP. We will analyse the reasons for the presence of a third party in health care, besides patients (the first party) and physicians (the second party). Then we will categorise the several types of relationships between both parties. The two research questions in this chapter are:

1. *What are the main functions of a third party in health care?*

2. *What are the implications of the main functions of a third party for the type of relationships between third parties and general practitioners?*

In chapter 3 we will introduce a theory that focuses on the relationships between two parties with asymmetric information and conflicting interests, and that proposes strategies to deal with the problems that may occur within such relationships: the theory of agency. We will use this theory later on to provide managed care with a theoretical background. Hence central to this chapter is the third research question:

3. *What are the main characteristics of agency theory?*

Whether agency theory is indeed suitable to analyse the relationships between third parties and GPs is subject of chapter 4. The research question is:

4. *Is agency theory applicable to relationships in the health care sector in general, and to the relationships between third-party agents and general practitioners in particular?*

Third parties can use a set of techniques in order to influence the behaviour of the GPs. These techniques will be analysed in chapter 5. The research question is:

5. *Which techniques can and do third-party agents apply within their relationships with general practitioners in order to reduce the agency problems within the patient-physician relationship?*

One of the potential techniques a third party may use is financial-risk sharing with the GP. The research question central to chapter 6 then is:

6. *What is the rationale for financial-risk sharing between third-party agents and general practitioners?*

Given that there is a rationale for financial-risk sharing, the question is how a contract can be devised that arranges the financial and the organisational aspects of the relationship. We will develop an analytical framework to compare the features and effectiveness of different systems of financial-risk sharing. This brings us to the second research question of this chapter, namely:

7. *How can systems of financial-risk sharing be structured?*

In chapter 7 we will review several examples of financial-risk sharing. Further, we will analyse whether the analytical framework of risk sharing created in chapter 6 is useful in analysing relationships in which the risk is shared. The two research questions are:

8. *What are actual effects of different systems of financial-risk sharing on the performance of general practitioners?*

9. *Does the analytical framework of financial-risk sharing sufficiently provide insight into the key differences of systems in which the risk is shared between third party and general practitioner so as to infer the effectiveness of such systems?*

The final research question then is:

10. *How should systems of financial-risk sharing be structured?*

We will end with a summary and conclusion.

2 THIRD-PARTY AGENTS AND GENERAL PRACTITIONERS

2.1 Introduction

As is in other markets, in health care one can discern two parties. The first party is the demander of care, like a consumer or a patient. The second party is the supplier of care: an individual health-care provider (like a physiotherapist, a dentist or a physician) or an institution (like a hospital or a nursing home). Besides these two directly involved parties, however, there is a role for a third party. A third party in health care can perform different functions, like taking over the financial risk from consumers or guarding the financial accessibility of health care (Van de Ven et al. 1994). In contrast to what the term 'third party' suggests, mostly it is not just one party that is involved. Rather, one has to think of various functions that can be performed by different governmental agencies and public or private organisations.
The first research question central to this chapter is:

What are the main functions of a third party in health care?

In section 2.2, the reasons for the presence of a third party and the different third-party functions are addressed. In section 2.3, the second party central to this research (the general practitioner) is introduced, and the term 'third party' is further defined. Also, the main types of relationship between third parties and general practitioners are described by means of several classification schemes.

2.2 Third-party functions

2.2.1 Characteristics of health care

An important characteristic of health care is the presence of *uncertainty* at the demand side (Arrow 1963). The uncertainty concerns the moment of consumption, the extent of consumption, and the effect of consumption. Because of the uncertainty about time and extent of use, there is a financial risk to an individual. By paying taxes or an insurance premium to a risk-pooling third party (such as a government or a health-insurance company) an individual is able to protect him against the financial consequences of health-care consumption.

A peculiarity of health insurance (or other forms of risk pooling in health care) is the way in which payments are being made in case of health care consumption. Because of the difficulty of determining the amount of damage caused by illness, benefits are usually

paid in kind or on the basis of expenditures being made. So, instead of on the assessed value of the amount of loss, payments depend on expenditures being made to repair the loss. However, these expenditures only approximate the 'necessary' or 'appropriate' costs to cure the illnesses (Pauly 1986, 1988a). As an individual has some control over the probability of loss as well as the level of costs incurred, moral hazard may result. Moral hazard in health care refers to the inclination of an individual to consume more or more expensive care because of a reduction in the marginal costs of consumption due to the presence of health insurance. This phenomenon is called consumer-induced moral hazard.

Another important characteristic of health care is the *information asymmetry* between demanders and providers of care (Arrow 1963). Trained and skilled as they are, providers of care mostly are better informed about clinical pictures, possible treatments and the effects of treatments. The lack of information of consumers leads to a demand for diagnostic and therapeutic information to be used in decisions about future consumption (Pauly 1978). However, the supplier of this information is often a supplier of medical services as well. Especially in a fee-for-service setting and if fees are higher than the marginal costs of production, there is an incentive for providers to influence demand for their services by altering the demanders' perceptions of need. This is called supplier-induced demand (Evans 1974). Demand inducement as a consequence of the presence of health insurance is known as supplier-induced moral hazard.

The presence of *externalities* is being mentioned as a third characteristic of health care (Van de Ven et al. 1994). An individual can derive utility from the consumption of health care by others. This may, for instance, be out of concern with general welfare of other individuals (altruistic preferences) or for fear of contagious diseases (egoistic preferences).

Based on the characteristics of health care and health insurance, Van de Ven et al. (1994) distinguished three main functions of third parties. These are the insurance function, the agency function and the access function.

2.2.2 The insurance function

The occurrence of illness is largely a stochastic process. In general, it is unpredictable whether a specific individual will develop a certain disease and, if so, to what extent medical care will be needed. Hence it follows that there is a financial risk to that individual. If it is assumed that individuals are averse to financial risks and that they want to maximise the expected value of their utility, then it can be seen that they will prefer health insurance with premium m to no insurance with an expected income reduction of amount m (Arrow 1963, Pauly 1968). As the risk of a financial loss leads to a demand for health insurance, a first main role of a third party in health care is to provide insurance covering that risk. This function can be labelled as the pure insurance function of a third party, which consists mainly of pooling risks and paying claims (Pauly 1988b).[1]

[1] In this connection Enthoven (1994, p. 1415) speaks of 'remote third parties', which he describes as '(...) third parties that are payors only (...).'

Pooling an infinite number of independent risks reduces the variability round the average loss (the law of large numbers) and therefore the third party's risk to a minimum (Pauly 1988b). However, since the number of risks will not be infinite and some interdependence among the risks may occur, some risk will remain to the third party.[2] A loading fee on top of the actuarially fair premium will ease this financial risk and can compensate for administration costs besides (Arrow 1963). This extra premium will reduce income to a larger extent than amount *m*. In spite of a premium that is higher than the actuarially fair premium, an individual may still prefer to pay for health insurance. The individual will have this preference as long as the utility derived from income after insurance is higher than the expected utility of income in the uncertain situation without insurance (that is, if the individual is sufficiently risk averse).

A striking phenomenon in health insurance is moral hazard. The problem of moral hazard stems for the most part from the way in which benefits are being paid. Although the difficulty of observing the preventive measures taken by the insured contributes to the problem, the main cause seems to be the fact that it is hard to assess the damage caused by a certain illness (Pauly 1986). To assess this damage one should at the minimum be able to show the presence and the severity of an illness. As this is practically unfeasible, the assessment problem is in general evaded by paying in kind or by indemnifying the consumer by making payments that are based on factual costs. The effect, however, of this kind of payments is that it distorts incentives and decreases the welfare gained by insurance. In absence of cost-sharing, such payments reduce the marginal costs to the consumer at the moment of consumption to zero.[3] As Pauly (1968) notes, depending on the price elasticity of the demand for care, the reduced user price will increase the amount of care demanded. So, the extent of moral hazard depends on the price elasticity of demand and will be zero in case of perfectly inelastic demand.

Besides moral hazard initiated by a consumer, one can distinguish supplier-induced moral hazard. In this case the inclination of an individual to consume more, or more expensive care is caused by a provider persuading him that he is in need of that care. The provider induces the demand knowing that insurance will reimburse the costs. Supplier-induced moral hazard can thus be considered as a special form of supplier-induced demand (Schut 1995).

Risk pooling, i.e. performing the insurance function, is a way to reduce the financial risk to which individuals are subjected. But at the same time it gives rise to the financial risk of extra or extra costly health-care consumption by distorting incentives to demanders as well as to providers of care. Moreover, this overuse can lower the quality of care and may cause iatrogenic illnesses (Franks et al. 1992). Hence, another important function to be performed in health care is the reduction of moral hazard and the dissemination of information about the quality of certain providers as well as the appropriateness of medical care.

[2] An example of the interdependence among risks is the occurrence of contagious diseases.
[3] Other costs like time cost and inconvenience are not taken into consideration.

2.2.3 The agency function

2.2.3.1 Reducing moral hazard

To restrain the negative consequences of performing the insurance function, a third party has to do more than being a financier of care. To that end, it has to perform a function that Pauly (1988b, p. 237) labelled as 'cost containment', 'expenditure control' or 'limitation of moral hazard'. Hence a third party should not only operate on the health insurance market, but also on the health care delivery market. Pauly (1988b, p. 240) argued that it is important to note '(...) that the *reason* why the insurer finds it advantageous to interfere in transactions with providers springs largely from the distorted incentives offered by insurance'.[4] This interfering in the delivery of health care is not easy. As Luft noted, 'it is one thing to be an efficient marketer of coverage and processor of claims and quite another thing to be a manager of a medical care delivery system' (Luft 1985). Van de Ven et al. (1994) spoke of the agency function of a third party in this connection. Their concept of this agency function is somewhat broader than the function described by Pauly. It consists of limiting moral hazard, buying care for a certain population, and collecting and providing information about (providers of) care. The agency function extends beyond just cost containment to acting as an agent for the consumers, for instance by being concerned about the appropriateness of care.

The moral-hazard problem can be divided in consumer-induced moral hazard and supplier-induced moral hazard. Third parties have several instruments at their disposal to reduce both problems. Constraining the insurance benefits, for example by establishing upper limits on the height of payments, may reduce consumer-induced moral hazard. These limits can be conditional on the disease or treatment ('quasi-indemnities') as well as unconditional (Pauly 1986). Other well-known methods of cost-sharing are co-payments, deductibles or coinsurance. Another way of constraining benefits is restricting insurance coverage to necessary care for which demand is highly inelastic.[5] Although these methods might influence provider behaviour, they are primarily directed at consumers of care and are beyond the scope of this thesis.

The (use of the) set of techniques that the third party has at its disposal to reduce supplier-induced moral hazard is often designated as 'managed care'. By means of one or more of such techniques, the third party may attempt to influence the decision-making process within the relationship between a patient and a provider of care. Examples of these techniques are (financial) incentives, utilisation review, mandatory second opinions or physician profiling. The main managed-care techniques are described in chapter 5.

Purchasing health care goods and services on behalf of consumers is an important part of the agency function. Health care purchasing means that remote third-party payers have

[4] Although Pauly confines himself to health insurers, the same seems to hold for other third parties within the health-care sector (like a government).

[5] A way of restricting the benefits package has been proposed by the Dutch Committee on Choices in Health Care (commissie Keuzen in de zorg). The committee proposed to screen benefits on necessity, effectiveness in relation to medical indication, and efficiency, and to assess whether consumers should account for the costs on their own (commissie Keuzen in de zorg 1991).

to abolish the model in which they simply indemnify consumers of care. There are, however, different ways in which this part of the agency function can be fulfilled. Purchasing activities range from simple agreements about the nature and price of services to advanced agreements about co-operation in the development and provision of services. A way of classifying the different arrangements is by distinguishing between 'contracting', 'purchasing' and 'commissioning', as proposed by Øvretveit (1995, p. 18). The narrowest concept is that of *contracting*, which is about selection and remuneration of providers and about specification of the nature of services to be provided. Like in the 'remote third-party payment' model, the third party has no intention to reduce moral hazard and to improve the consumers' health status. An important difference is that by contracting providers, the third party arranges access for consumers to the contracted services.

Health care *purchasing* is a broader concept and can be defined as 'buying the best value for money services to achieve the maximum health gain for those most in need'. Besides contracting activities, it consists of the assessment of needs, planning of required services, deciding on a purchasing strategy, handling of complaints et cetera. Moreover, the purchaser pays attention to the efficiency – technical as well as allocative – and the effectiveness of the services being bought.

The broadest concept, however, is that of *commissioning*. According to Øvretveit, the purpose of health commissioning is 'to maximise the health of a population and minimise illness by purchasing health services and by influencing other organisations to create conditions which enhance people's health'. Commissioning is more than purchasing services: it also encompasses, for instance, the stimulation of providers to plan, establish and co-ordinate services so that these services can be purchased when they are needed, and the co-ordination with other purchasers of the planning and contracting activities. Although health commissioning stresses on health care, commissioning third parties may also put means into other health-influencing factors, like housing or working conditions.

2.2.3.2 Past performance of the agency function

Besides being beneficial to consumers, the performance of (parts of) the agency function, especially reducing moral hazard, seems to be in the interest of third parties themselves. Nevertheless, many third parties have been reluctant or unable to perform the agency function (Van de Ven et al. 1994). This reluctance or inability differs per country, and even differs per third party within a country though. In the United States (US), for example, federal and state governments have been much more willing to perform parts of the agency function than other third parties, like private purchasers (IOM 1989).

With regard to conventional private health insurance, Schut (1995) mentioned several reasons for the prominence of a model in which the insurance function prevails. Firstly, during the rise of private health insurance, providers of care were able to further a model in which there is no relationship between third parties and providers and in which consumers are reimbursed for health care expenses. Following from this is that without a contractual relationship with providers – and especially in case of insurers having small market shares – it has been difficult to perform the agency function. Thirdly, such a model with an insurer remotely situated entails a free choice of providers, which is attractive to consumers as well as providers of care. The fourth reason is the interest insurers

collectively have in rising health care costs. A larger financial risk increases, to a certain extent, the demand for health insurance. A last reason is that, on the one hand, an individual insurer may have an incentive to try to curb rising premiums by performing the agency function. But, on the other hand, this creates a free-rider problem since other insurers may profit from measurers taken by the first insurer.

Not just private health insurers, but also several public health insurers (like the former Dutch 'sickness funds') as well as several governmental pooling systems (like the British National Health Service) have been unwilling or unable to perform the agency function properly. Dutch public insurers were lacking the incentive to reduce inefficiencies in health care because of the retrospective payment of all costs by the General Fund – a system that was in force from the Second World War until the nineties. As they were (until 1992) obliged to contract with every provider of care within their area, public insurers were not able to buy care from selected providers (Schut 1995). Hence there was practically no (financial) incentive to reduce moral hazard and it was legally impossible to be a careful and prudent buyer of care. There was no incentive to court the favour of insured either, because insured were assigned to the insurer in their area. During the last decades, the Dutch government performed the cost-containment function by constraining supply and prices.

The same kinds of arguments seem to be applicable to a National Health Service, such as the old-style NHS in the United Kingdom (UK). Until the reforms – these were introduced in 1989 and implemented from 1991 on – everyone was assigned to a District Health Authority that arranged all hospital and community care. In such a largely integrated and monopolistic system, incentives to strive for efficiency or for responsiveness to consumers' needs or wishes are limited. Although there was a pressure to contain costs in the NHS, it was not so much a pressure to act as the consumers' agent by reducing moral hazard or organising effective and efficient care. Rather, it was a pressure coming mainly from the top (the NHS Executive) in an attempt to keep within the expenditure limits set by the Treasury (Propper 1995a).

2.2.3.3 Increasing importance of the agency function

There are several reasons why the importance of the third party's agency function has been increasing or will increase in the near future. A first reason is the reasonable expectation that there will be a further rise in health care expenditures. Newhouse (1992) mentioned a number of factors that have caused such an increase and that might cause a further increase. Firstly, and according to Newhouse perhaps most importantly, health care expenditures will rise as a result of technological change, or what he called '(...) the march of science and the increased capabilities of medicine' (Newhouse 1992, p. 11). Secondly, health care is a labour-intensive service in which it is difficult to raise productivity. Although some productivity gains are possible – reduced length of stay in hospitals is an example – on average these gains will be lower than in other sectors of the economy. However, if wages in the health care sector follow wage increases in other sectors, health care will become relatively more expensive. If volume of care remains the same, quality of care will tend to go down or total expenditure will tend to go up (Baumol's cost disease of the service sector). Thirdly, the increase can be explained by the ageing of

populations. People live longer and, therefore, are more likely to have on average higher health care costs. Moreover, older people are also more likely to develop a costly chronic disease.

Another reason why the third party's agency function is becoming or will become more prominent, is a changing opinion about the consumer's position in health care. As is the case with other goods and services, consumers want more freedom of choice and responsiveness to their needs and wishes (Glennerster et al. 1994). To make choices, sufficient information is needed. Since insurance payments reduce marginal costs at the moment of consumption to zero, it is information about the quality of care that is relevant to consumers, not about prices. In general, consumers have difficulties in judging the skills of health care providers and the quality of diagnosis and treatment. Third parties are more able to collect and to analyse data, and to disseminate information about practice patterns, quality and outcomes of health care than individual consumers are. As far as such data has been collected in the past, it has not been disseminated because of imperfections in the data or because it was meant for internal use only (Pauly 1988a).

2.2.4 The access function

As argued, performance of the insurance function and the agency function by a third party may overcome problems created by two of the intrinsic characteristics of health care, namely uncertainty and the asymmetry of information. The third characteristic, external effects, also legitimates third-party involvement in the health-care sector though. Since whether or not consuming health care by a certain individual may have consequences for the physical or psychological well-being of others, there is a societal concern over the accessibility of at least some basic goods and services.

Externalities in health care stem from different interpersonal preferences. Firstly, in case of *altruistic preferences*, a certain person derives utility from another person's well-being. This other person's well-being is a function of his health – a status that may be influenced by health care consumption – and his consumption of other goods and services. Altruistic preferences can give rise to an income-transferring scheme in order to enable the benefiting person to rise his utility. However, the person who obtains additional means decides whether or not to spend these cross-subsidies on health care.

More specific are the *paternalistic preferences*. One might particularly be concerned about someone's health status or about his use of health care, which influences this status. Such preferences can lead to a demand for third-party interventions in order to improve the financial accessibility of care. Possibilities are subsidising premiums, subsidising individual or institutional health care providers, or supplying health care.

In case of contagious diseases, *egoistic preferences* with regard to accessible health care may result in subsidisation or free provision of preventive and curative services.

It's true that performance of the insurance function may remove economic barriers to access, but this is not necessarily the case. Absent a compulsory risk-pooling arrangement, some individuals might prefer not to insure, thereby risking the inability to pay if health care is needed. Furthermore, in an unregulated, competitive insurance market

some individuals may encounter serious problems in obtaining insurance. As in such a market there is a strong tendency towards experience rating, some high-risk individuals will be unable to pay a premium that reflects their risk. Problems in obtaining insurance may also result from a failure in the health-insurance market: for fear of adverse selection, the risk-pooling party may refuse to accept the risk or offer coverage at a premium that is, in view of the individual's risk, mistakenly too high (Evans 1984).[6]

Although it is an important function, in this thesis the access function is left out of consideration.

2.3 Relationships between third-party agents and general practitioners

2.3.1 Introduction

Third-party payers and third-party agents

Like the insurance function and the access functions, the agency function may be fulfilled by a third party. In this thesis a third party is considered to be a *risk-bearing party* receiving contributions for a *defined group of members* and for a *certain period of time*, and having the *contractual or legal obligation to reimburse or to provide a specified package of health-care goods and services*. An example of a third party covered by this definition is a health insurer. Notice that this definition does, for instance, not comprise governmental agencies that merely develop legislation to protect patients or that merely fulfil the access function, or a Central Fund that merely collects premiums and makes payments to health insurers.

Then, there is the distinction between third-party payers and third-party agents. We suppose that a *third-party payer* is mainly concerned with performing the insurance function and has no intention to arrange access for the insured to health-care services and to improve the insured' health status. Probably, emphasis will be placed on marketing and selling of insurance products, claims processing, and the like. A *third-party agent*, however, is supposed to act on behalf of the insured or patients. This acting may manifest itself in, for instance, reducing moral hazard by using a wide array of measures to influence the way health care is provided.

General practitioners

The second party in this thesis is the general practitioner. The term general practitioner is used here without reference to a specific kind of doctor practising in a specific health-care system. The term primary care physician is also used in case we referred to the United States or in case we cited from articles in which the term is used. Although primary care physicians and general practitioners (GPs) may differ from each other in some aspects, they have some common characteristics – or at least, they ought to, otherwise

[6] Adverse selection arises from the asymmetry of information between a buyer and a seller of insurance. The less-informed seller of an insurance policy runs the risk of attracting buyers who expect their individual loss to exceed the premium they have to pay.

they can not be distinguished from other (specialty) physicians (Starfield 1992). These characteristics are (Starfield 1992, Fry and Horder 1994):

- The provision of direct accessible first-contact care to a small, defined population.
- Longitudinal care, which means that '(...) individuals in the population identify with a source of care as ''theirs,'' that the provider or groups of providers at least implicitly recognise the existence of a formal or informal contract to be the ''regular source of care,'' and that this relationship exists for a defined period of time or indefinitely until explicitly changed' (Starfield 1992, p. 41).
- The care provided or arranged for is comprehensive and consists of all types of goods and services, including referrals.
- There is co-ordination (integration) of care, for instance by means of a regular physician and medical-record keeping. The physician has information about the patient's prior problems and the goods or services provided before and recognises the meaning of that information for present actions.

This is more or less comparable to the attributes of primary care summarised by Godber et al.: direct access, generalist care, longitudinal care and delivery in a community setting (1997, p. 276).

There were two reasons for our choice of the GP as the second party. Firstly (following Starfield 1992, p. viii), the choice of physicians is because of the overall responsibility they usually bear. Nurses, for instance, may bear responsibility for some parts of primary care too, but are usually not responsible for the totality of the care. Secondly, from a third-party agent's viewpoint, the relationship with a GP may be interesting as these physicians often fulfil a role as gatekeeper and co-ordinator of medical services. As a result, these physicians have a considerable influence on the nature, quantity and quality of the goods and services delivered.

2.3.2 Classification of relationships between third parties and general practitioners

The relationship between a third-party agent and a GP can be structured in various ways. Question then is whether the structure of such relationships should be contingent on the third-party functions. In order to answer the second research question,

What are the implications of the main functions of a third party for the relationships between third parties and general practitioners?

we will first take a closer look at the relationships by reviewing several classification schemes.

The classification of Hurst

Hurst (1992) distinguised between three models. The *reimbursement model* is characterised by the absence of a relationship between second and third party. The third party only fulfils the insurance function. It pools risks and pays claims (i.e. reimbursing the patient for incurred health-care expenses) and has no contractual or legal obligation to arrange or to provide health care goods and services. In the *contract model* there is a contractual

relationship between independent parties. At least arrangements are made with regard to the accessibility of the contracted services. Often, payments are made directly to the physicians so that health care is provided in kind.[7] The *integrated model* concerns a vertically integrated system in which health insurance and care are supplied by the same organisation.

The ability to perform the agency function will differ per type of relationship. In absence of a relationship it is virtually impossible to perform the agency function. A contractual model is more suited then as both parties may agree upon a contract that embodies, for instance, the third party's interference in the provision of health care. The integrated model may seem even more suited because of the complete integration with an employer-employee relationship and the resulting instruments for the third party. Employment of the physician does not mean that he automatically acts in the best interests of both third party and patients though.

The classification of Van de Ven et al.

In case a third party not only fulfils a payer function but also acts as an agent on behalf of the insured, another classification might be more suited. Van de Ven et al. expanded Hurst's framework by combining it with four alternative health care markets. The latter resulted from the distinction between competing and monopolistic third parties on the one side and competing and monopolistic health care providers on the other.[8] This combination yields ten models of third party-provider relationships.[9] Expansion with market structure is valuable since this structure provides third parties and providers with incentives that may influence behaviour. For instance, a competitive market may provide a third party with an incentive to perform the agency function properly. In a monopolistic market this incentive is absent. On the other hand, without competition it may be easier to perform this function, as a monopoly provides the third party with the necessary leverage to manage care (Van de Ven et al. 1994).

Their classification is also useful as it touches on the position of third party and physician in contract negotiations by distinguishing the structure of both sides of the market. Both parties have bounded possibilities to offer and to refuse contracts depending on the market structure. In case of a monopolistic contract model, for instance, third parties may be limited in their contract design. This will be different in a monopsonistic contract model. In case of competing third parties, physicians may have several options to choose from. In a monopsonistic contract model, however, they may be entirely dependent upon the contracts the third party offers.

[7] Notice that Øvretveit's contracting concept, as described in subsection 2.2.3.1, is narrower and only one possible form of Hurst's contract model.

[8] In this classification, only price competition is taken into consideration.

[9] Combinations with an integrated model only exist in a situation of competition between third parties *and* competition between providers, or in a situation without competition between third parties *and* without competition between providers.

The classification of Weiner and de Lissovoy

Weiner and de Lissovoy (1993) described a model in which the contractual relationships between consumers, sponsors, providers and intermediaries are identified.[10] They defined six dimensions of the contracting process: whether the sponsor, the intermediary or the physicians assume a financial risk, whether the consumer is free to select a provider, whether the provider's autonomy is restricted, and whether the health plan is obliged to provide care. Of interest here are the placement of risk on the physicians and the restrictions on their practice choices. Whether for cost-containment or for quality reasons, the contract is used to influence the way health care is provided.

Like Hurst, Weiner and de Lissovoy (1993, pp. 89-90) described an integrated model in their classification scheme. In their scheme, an integrated system is a health plan where:

1. There is a legal responsibility to deliver medical services to enrolled consumers who seek care from within an integrated network of providers employed by, *or under contract to*, the plan (italics added, A.V.).
2. There is an entity that manages care by controlling the patterns of practice of providers in the network. This is accomplished by administrative and possibly financial controls. These include, at a minimum, mandated pre-certification of major services and retrospective profiling of provider practices via information systems.

In Hurst's classification the integrated model is characterised by the presence of labour contracts, whereas the definition of an integrated system as proposed by Weiner and de Lissovoy is broader and more functionally oriented. The implication is that an arrangement that is of the integrated type according to Weiner and de Lissovoy, may resemble Hurst's contract model.

The classification of Welch et al. and Hillman et al.

A classification scheme can also highlight the financial as well as the organisational side of the arrangements. Welch et al. (1990) and Hillman et al. (1992) developed such a classification scheme.[11] Their framework is specifically aimed at the relationship between third parties and primary care physicians. Although they focused on Health Maintenance Organisations (HMOs), their scheme seems equally applicable to relationships in which other parties fulfil the insurance function and the agency function.[12] Therefore, a modified version is discussed here: it is made applicable to relationships with other third parties, and the order of the characteristics has been changed. The following characteristics are discerned:

1. Whether the third party contracts the primary care physicians directly (two-tiered system) or indirectly via an intermediary organisation (three-tiered system).

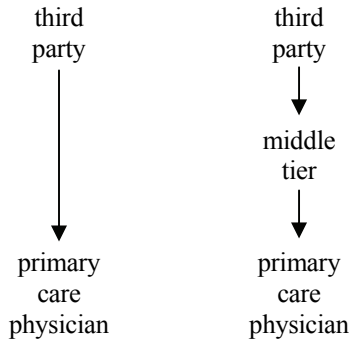
[10] In their taxonomy sponsors, like employers or the government, pay the majority of costs of a health plan. Intermediaries may have administrative functions only, but may also co-ordinate care or bear the insurance risk.

[11] The difference between the classifications proposed in both articles is the order in which the characteristics are arranged.

[12] An HMO is an organisation with a contractual responsibility to arrange and to provide health-care services to a population of enrolled subscribers (see subsection 7.4.1).

2. The method of payment by the third party, including:
   a. the basic method of payment for primary care services;
   b. the size and nature of the ancillary payments, if any.
3. The way an intermediary organisation, if any, translates the contracts with the third party into contracts with the physicians.
4. The size and nature of the risk pool, if any, used to share the risk or reward.
5. Whether the primary care physicians see only members of this third party or also of other third parties.

Ad 1
In a two-tiered system there is a direct relationship between the third party and the physician, whereas in a three-tiered system an intermediary organisation (the middle tier) is located between the third party and the physician.



Two-tiered system Three-tiered system

It is important to distinguish between these two organisational forms. A middle tier or intermediate entity may change the incentive system (or other parts of the arrangements) the third party uses. Further, the nature of the relationship between third party and physician may differ from the one between middle tier and physician (Welch et al. 1990). This will especially be the case if the middle tier is a group practice consisting of colleagues practising within the same building. Gold et al. (2002) pointed out that the concept of two- and three-tiered systems is even outdated and oversimplified since complex arrangements can lead to four, five or even more tiers (multi-tiered arrangements). Further, they pointed at the fact that risk sharing with an intermediate entity is often combined with the delegation of responsibilities, for instance for provider selection.

Ad 2
There are several methods for paying physicians, like capitation or fee for service. The incentives the physicians experience differ per method of payment, so the third party may use them as a first instrument to influence the physician. Modifying the nature and the extent of ancillary payments is another way of influencing the physician. Ancillary payments may, for instance, take the form of bonuses.

Ad 3
The contract between third party and middle tier may be translated into a contract with different incentive mechanisms. For example, the third party may use capitation to pay a directly contracted physician. If it pays a middle tier according to a capitation scheme, the middle tier on its turn may pay the physician by means of fee for service. The middle tier will be motivated, though, to 'pass through' the incentives it faces (Eggleston 2005). The example of a capitation scheme that is translated into a fee-for-service system may result in financial problems for the middle tier.

Ad 4
The fourth characteristic is the risk pool. A risk pool is a group of providers who share in the rewards and penalties from surpluses and deficits in, for instance, drugs or referral budgets.

Ad 5
Whether or not physicians also see patients who are member from another third party, is determining the amount of freedom patients have in choosing among physicians and third parties. More important here, however, is that the proportion of patients who are member of a certain third party is determining the responsiveness of the physician to the financial incentives employed by that third party (Welch 1990). It may be expected that the larger this proportion is, the more the physician will be responsive to the incentives.

2.4 Conclusion

In this chapter we addressed some issues with regard to the functions of third parties in health care. To find an answer to the first research question,

What are the main functions of a third party in health care?

we first described some characteristics of health care. The sector is characterised by uncertainty at the demand side, an asymmetry of information between demanders and providers of care, and the presence of external effects. These characteristics form a justification for the presence of a third party. Three functions of a third party in health care are:
- the insurance function, which consists mainly of pooling risks and paying claims;
- the agency function, which consists of limiting moral hazard (for instance by buying care for a certain population) and collecting and providing information about care;
- the access function, which is about guaranteeing the accessibility of, at least, some basic goods and services.

In this thesis, the third party is considered to be a *risk-bearing party* receiving contributions for a *defined group of members* and for a *certain period of time*, and having the *contractual or legal obligation to reimburse or to provide a specified package of healthcare goods and services*. As second party is chosen for the group of GPs because of the

characteristics of GP care and the GP's role in health care. From the viewpoint of a third-party agent, the relationship with GPs may be of interest as they often act as gatekeepers and co-ordinators of medical services. Hence GPs have a considerable influence on the nature, quantity and quality of health care.

Third parties that act as agent on behalf of the insured have to enter into relationships with providers of care, among whom GPs, in order to fulfil their role properly. Fulfilling the agency function is difficult in case a relationship between both parties is absent, as in a reimbursement model. The relationships may be structured in different ways, ranging from simple contractual arrangements to full integration. This brings us to the second research question central to this chapter:

What are the implications of the main functions of a third party for the type of relationships between third parties and general practitioners?

There are several ways the relationships may be analysed and classified. We reviewed classification schemes of Hurst, Van de Ven et al., Weiner and de Lissovoy, and of Welch et al. and Hillman et al. to provide some basic insight into the various types of relationships. The classification of Hurst is of interest because it gives insight into the main financial and contractual relationships between third party and physician. The classification of Van de Ven et al. adds to this the market structure, which is an important factor for the possibilities and the incentives to perform the agency function. The classification of Weiner and de Lissovoy adds the concepts of the distribution of the financial risk among several parties and of restrictions on the physicians' clinical options. Finally, the classification of Welch et al. and Hillman et al. is specifically aimed at the relationship between third parties and primary care physicians and analyses the financial and the organisational side of the arrangements between both parties in detail.

Two major conclusions can be drawn from the four classification schemes. A first conclusion is that a particular relationship between a third party and a (primary care) physician can be approached in several ways, like in a juridical, an organisational or a financial way. But even then the several authors highlight different aspects of, for instance, the organisational side of relationships. Not all of them included the option of the middle tier in their schemes, for instance. If one is interested in the payment system used by a third party, the presence of a middle tier may not be relevant. But if one is interested in the effect of that payment system on the behaviour of GPs, the recognition of a middle tier, which may alter such a payment system, is crucial.

A second conclusion is that the financial as well as the organisational structures of relationships between third parties and GPs can differ considerably. These may vary from no relationship at all to full integration of both parties. Third parties may contract directly or indirectly (via a middle tier), pay physicians directly or reimburse the insured, use several payment systems, et cetera. Middle tiers may alter a contract concluded with a third party and change, for instance, the payment system. Such aspects of the relationships and several of the concepts of the classification schemes will be dealt with in following chapters.

3 AGENCY THEORY

3.1 The choice of a theory

3.1.1 Introduction

In the previous chapter, the two parties central to this thesis were introduced. Firstly, we introduced the concept of the third-party agent: a risk-bearing party receiving contributions for a defined group of members and for a certain period of time, and having the contractual or legal obligation to reimburse or to provide a specified package of healthcare goods and services, and that is supposed to act on behalf of the insured or patients. Secondly, we introduced the general practitioner (GP): a direct accessible, general physician who provides longitudinal, comprehensive care in a community setting. Acting on behalf of the insured patients, the third-party agent may want the GP to provide care in a way that is beneficial to the patients. There are several ways the third-party agent can attempt to influence the behaviour of the GP and to promote a certain outcome. In order to achieve that, the third-party agent may want to enter into a contractual relationship with the GP. In this chapter we will discuss the (choice of a) theoretical background for our study of the various relationships between third-parties and GPs and of the various strategies the third party may use.

In section 3.2, we will explain the theory itself. In section 3.3, we will present several means to handle agency problems. Subject of section 3.4, finally, are the theory's strong and weak sides, and the advantages and disadvantages of using the theory in this thesis. We will end with a summary. The application of the theory to the health care sector and, more specifically, to the relationship between third-party agents and GPs, is subject of the next chapter.

3.1.2 Theoretical framework

In order to exert influence on the way care is provided, the third-party agent may want to enter into a contractual relationship with the GP. In contract theory is distinguished between complete contracts and incomplete contracts (Hendrikse 2003).[1] Complete contracts are considered to be complete because they contain all the information that is available to the contracting parties. Not all of the information may be available to both parties, but as far as it is and as far as it is relevant it will be part of the contract. This

[1] The paragraph on contracting theory is largely based on Hendrikse (2003).

information pertains to the actions of the contracting parties and the possible future situations. Everything that is known to the contracting parties is specified ex ante, so ex-post bargaining because of unforeseen situations is not needed. As it is generally assumed that parties show rational behaviour and that all the relevant information can be included freely in the contract, complete contracts may become very complex.

In complete contract theory no distinction is made between observable actions and verifiable actions. Observable actions are actions that are observable by the parties involved in the contract. These actions are termed verifiable, if they can be verified by a third party. This may be important in case of conflicts. In complete contract theory it is assumed that all observable information is also verifiable and that potential conflicts are dealt with by means of a contract – opportunistic behaviour will be taken account for during the design of the contract – or a third party (like a judge).

An influential theory that represents complete contracting theory is the classical agency theory. This theory focuses on the contractual relationship between two parties. One party (called the agent) acts for or on behalf of the other party (called the principal). Their relationship is characterised by conflicting interests and asymmetric information. The theory stresses the issues of compensation based on measured performance and monitoring. The principal has to devise an incentive scheme in such a way that the agent is stimulated to act in the principal's interests, given the behavioural assumptions of rationality and opportunism. Agency theory was initially developed to investigate questions of incomplete information and risk sharing. Later it became an analytical tool for organisational theory as well (Moe 1984). Hereafter, we will return to this theory extensively.

Complete contract theory has been criticised for several reasons. This has led to the development of alternative theories, which can be designated as incomplete contract theories. There are several reasons for contracts to be incomplete (Hendrikse 2003). Firstly, it may not be attainable to anticipate all possible contingencies. Hence incomplete contract theory disputes the assumption of complete contract theory that it is possible to foresee every future action and every possible future situation and to include these possibilities into the contract for free. Secondly, complete contracts may become complex, and writing such a contract may be too costly. Thirdly, language may be context-dependent, allowing for divergent interpretations of the terms of the contract. Fourthly, not all the relevant information may be verifiable. Incomplete contracts will only contain observable actions that can be verified. Given the behavioural assumption of opportunism and in the absence of trust, it makes not much sense to agree upon non-verifiable actions. In case of a conflict then, the cause of the conflict can not be proven to a third party (like a judge).

Complete contract theory has also been criticised because it can not explain the boundaries of organisations. In fact, in complete contract theory organisations are not very relevant as everything is covered by the contract. There is not really a difference between contracts within or contracts between organisations, although organisational aspects may influence the incentives parties face. In complete contract theory, there are solely ex-ante decisions. In incomplete contract theory, however, it is assumed that the

incomplete contract does not account for all possible contingencies. As a result, a situation may occur for which the contracting parties made no provisions in the contract. Hence there is a difference between ex-ante decisions and ex-post decisions. Given the behavioural assumption of opportunism, this may create ex-post problems like violation of the contract or renegotiation over contract conditions. In dealing with this potential ex-post opportunistic behaviour, a governance structure, like an organisation, can play an important role. Incomplete contract theory tries to analyse and to explain why particular governance structures are or should be chosen in case of specific circumstances.

Several attempts have been made to explain why in some circumstances the organisation instead of the market is chosen as the governance structure. Alchian and Demsetz (1972), for instance, viewed the integrated firm as a means to increase production through team production. They considered to some extent teamwork within a firm to be more productive than production through the market, because of problems with the monitoring of the agents' efforts. They argued that monitoring can be done better within a firm and that that it will be done appropriately if the party that monitors receives the residual income resulting from the team's actions. Monitoring also played a role in the analysis of Holmstrom and Milgrom (1991), who argued that the costs of measuring the agent's performance are an important determinant of integration. In situations in which the costs of measuring the agent's performance are low, it may be favoured that the agent owns the assets. If the costs are high, integration may be favoured.

An influential approach to the problem of choosing a governance structure has been the theory of transaction costs economics. It can be viewed as an economic theory of organisation. It focuses on the choice of a certain governance structure, given several characteristics of a transaction. Two important governance structures are the hierarchical organisation and the market. Question then is why some transactions take place within the hierarchical organisation and other transactions take place on the market. According to Williamson (1985) environmental factors, like uncertainty and small-numbers exchange, and human factors, like bounded rationality and opportunistic behaviour of agents, are important aspects for the relative efficiency of governance structures. A crucial factor is asset specificity, which refers to the relative lack of transferability of assets that are used in a given relation to other relations. Examples of types of asset specificity are site specificity or physical asset specificity. Especially if asset specificity and the degree of uncertainty are high, and given that rationality is bounded and agents are opportunistic, then internalisation, or at least very strong contracts, will be optimal. The hierarchical governance structure is supposed to prevent ex-post bargaining in case an unforeseen situation emerges. The thread of ex-post bargaining may lead to a situation in which parties don't want to invest ex ante in the relation-specific assets. As a result, the transaction may not take place at all. If the hierarchical governance structure takes away this thread, parties may be willing to make the necessary investments.

One reason that transaction costs economics has been criticised is that it argues what the costs can be of organising a transaction on the market, but that it does not argue what the costs can be of a transaction within a firm. The theory does for instance not explain why opportunistic behaviour of parties changes within an organisation, and it does not

explain either why a firm would stop integrating if integration reduces transaction costs. It also does not give a clear definition of integration (Grossman and Hart 1986). According to Grossman and Hart, a governance structure can define and allocate the decision rights for those situations for which no provisions were made in the contract: the residual decision rights. They distinguished between residual rights and specific rights. Their theory of vertical and lateral integration focuses on the ownership of assets. They presented it as a theory of costly contracts that emphasizes that contractual rights can be divided into specific rights and residual rights. As contracting is costly, it may be too costly to specify all the rights a party may want to have over another party's assets. Then, it may be optimal to specify some specific rights of control over the assets in the contract and to purchase all the other rights (the residual rights). Ownership then, implies the right to control all the aspects of the assets (the residual rights) except those that have been explicitly given to the other party by means of a contract (the specific rights). Grossman and Hart argued that when the residual decision rights are purchased by one party, they are lost by another party. This creates distortions, and therefore, they argued, there is also a cost to integration. 'That is, integration shifts the incentives for opportunistic and distortionary behavior, but it does not remove these incentives' (Grossman and Hart, 1986, p. 716). The approach of Grossman and Hart (1986) and of Hart and Moore (1990) is a property rights approach whereby the point of view is taken that for decisions upon integration, the possession of the rights of control over particular assets is crucial. In other words, if one party wants to have the decision rights concerning another party's assets, it needs to integrate.

The relationship between parties in which transactions take place can thus vary by choice of a governance structure. Given that contracts are incomplete and ex-post bargaining problems may occur, a governance structure can determine the way these ex-post bargaining problems are handled, for instance by giving a party the authority to decide in circumstances not covered by the contract. Hendrikse and Jiang (2005) argued that a governance structure is concerned with two important questions.[2] One important question is how the decision rights are allocated and who has the right (in the form of authority and responsibility) to take decisions regarding the deployment and use of assets. Themes with regard to authority are, for instance, the allocation of authority, formal versus real authority, decision control (ratification, monitoring), decision management (initiation, implementation) or enforcement mechanisms. A second important question is how the benefits and costs associated with the use of an asset are allocated. Themes with regard to income rights are, for instance, the use of financial incentives and compensation schemes, and who receives the benefits and who has to pay the costs of using an asset.

In the above, the agency approach is referred to as an example of a complete contract theory. Agency theory is concerned with the analysis of income rights. As an incomplete

[2] The distinction between decision rights and income rights was made by Hansmann, H. (1996), The Ownership of Enterprise, The Belknap Press of Harvard University Press, Cambridge.

contract theory, the property rights approach is concerned with the analysis of decision rights. This classification may be correct as far as it is related to the classical treatment of agency problems, concerned with complete contracts consisting of optimal compensation schemes. However, the distinction between the two approaches is becoming less strict. Holmstrom and Milgrom (1994), for instance, tried to integrate several approaches to the make-or-buy decision of firms. They argued that several choices regarding this decision are intertwined, like:

- the way agents are paid (for instance, a fixed wage or based on measured performance);
- the ownership of assets (owned by a firm or by an independent agent);
- the design of the job (whether the firm of the independent agent decides about tasks that are included or that are expressly excluded from the job, methods, or working hours).

They stated that these instruments can and should be used complementary. One of their findings was that if the costs of monitoring an agent are high or if some activities of the agent are important but hard to measure, then it is more likely that commissions are modest, the firm owns the assets, and job restrictions limit the agent's freedom. On the other hand, if monitoring an agent is easy or if there are no important hard-to-measure activities of the agent, then it is more likely that commissions are strong and output-based, the agent owns the assets, and the agent is free to design his own job.

Further, Neelen (1993) argued that in agency theory the problem of the principal is mainly confined to the design of an incentive structure, but that there are other control mechanisms that the principal may use to induce the agent to act in the principal's interests. One of these mechanisms is control by directives or authority, which is an important theme addressed by the decision rights approach.

We will analyse the relationships between third-party agents and GPs using agency theory, but we will continue Neelen's line of reasoning and expand upon the set of strategies that the principal may use to further the outcome he aims for. We will also address some modifications of, or extensions to the traditional assumptions of agency theory, which supports our view that the distinction between the agency approach and the property rights approach is becoming less strict.

One reason why we choose for an agency perspective and focus on concepts that are stressed by that theory is that it seems to correspond well with the situation central in this thesis. This situation relates to the contractual relationships between two parties (that is, third-party agents and GPs) in which the one party (the principal) enters into a relationship with a second party (the agent) in the expectation that the agent's actions are beneficial to the principal. As an important aspect of agency theory is the use of (financial) incentives, it corresponds well with the subject of this thesis, namely financial-risk sharing.

A focus on the allocation of decision rights (i.e. a property rights approach) could be interesting as well. An interesting question could for instance be why GP care is mostly provided by independent GPs, but sometimes by employed GPs. The question then could

be whether the authority of the employer changes the allocation of decision rights regarding the treatment of patients, or whether in case of employment the professional or the physician's individual autonomy is maintained or not. We assume though that, because of the professional or the physician's individual autonomy, authority within an employment relationship between third parties and GPs does not extend to interfering in the patient-physician relationship in a way different from the ways in a market relationship.

Another reason why we do not focus on the allocation of decision rights is that it concerns the ownership of or the contractual rights over specific assets. The use and the ownership of assets are less important issues in general practice than they are for instance in medical specialist care. Further, the assets related to providing GP care are hardly relation-specific assets. In other words, we consider the level of asset specificity to be low. In chapter 5, we will discuss instruments that in fact specify the rights over the use of, for instance, hospital assets. The GP owns most of the rights to decide whether or not to treat a patient or to have a patient treated, for instance by means of hospital assets. The third party has some specific rights, for instance, concerning the decisions whether or not to certify a hospital admission. Obviously, this departs from the relationship and ownership issues as described by Grossman and Hart (1986) because the GP's residual rights to decide whether or not to have a patient treated in hospital do not imply ownership of hospital assets. In fact, in the present relationships it is not so much a question of the rights over assets but of the rights over actions. Given the subject of financial-risk sharing, we are not interested in the organisational boundaries of firms and hence in the determinants of integration. We are interested in the contractual relationship between two parties. Moreover, we assume that, because of the professional relationship, integration does not solve the characteristic problems of agency relationships resulting from asymmetric information and conflicting interests. Partially analogous to the argument of Grossman and Hart (1986, p. 692), it is unclear how integration changes the scope for opportunistic behaviour 'when one of the self-interested owners becomes an equally self-interested employee of the other owner'. Pratt and Zeckhauser (1985, p. 32) argued that 'forming an organisation only internalises agency relationships. It does not eliminate problems of co-ordination, incentives, and so on. Presumably, though, it should substantially lessen them, for now we are all working on the same team and can be rewarded accordingly'.

The fact that we use agency theory as the theoretical background for our study does not imply that we will restrict ourselves to the standard economic treatment of the principal-agent problem with the issues of performance-based compensation and monitoring. We will argue hereafter that there are two approaches to agency problems, the normative, non-empirical principal-agent approach and the empirically based positive theory of agency. The positivist stream has been mainly concerned with describing the several governance mechanisms that should solve the agency problems (Eisenhardt 1989). Our approach is partly a positive one whereby we will describe mechanisms or techniques third parties may and actually do use within their relationships with GPs. Like Holmstrom and Milgrom (1991) and Neelen (1993), we will argue that the range of these

techniques is much wider than just deciding how to pay for performance. In chapter 5, we will try to bring these techniques in connection with each other. Partly, our approach is normative in the sense that we will discuss how systems of financial-risk sharing should be structured.

In this thesis, we will use agency theory to examine different financial and organisational arrangements between third parties and GPs as well as the strategies that third parties (may) use as part of their agency function.[3] Hence central to this chapter is the third research question:

What are the main characteristics of agency theory?

3.2 The theory of agency

3.2.1 Introduction

In this section we give an introduction to the economic theory of agency. Central to the theory are two phenomena that may jointly occur in relationships between two parties. These phenomena, which are asymmetrically distributed information and conflicting interests, are discussed in subsection 3.2.2. Although these phenomena were recognised earlier, the theory of agency was only developed in the seventies. It is one of the approaches within neo-institutional economics in which the applicability of the neo-classical theory of the firm has been broadened.[4] According to Jensen (1983, pp. 334-336), agency theory has been developed into '(...) two almost entirely separate and valuable literatures that nominally address the same problem.' One approach has been called 'principal-agent'. It is normative and non-empirical. Also, in general, it is highly mathematical. Authors concentrate on the most efficient contract given different levels of information, risk aversion and uncertainty.

Another approach is what has been called the 'positive theory of agency'. In this type of literature is also focused on the relationships between principals and agents, but contrary to the principal-agent approach it is, in general, non-mathematical and empirically oriented. Positive agency theory is '(...) about how the world behaves' (Jensen 1983, p. 320). The theory is aimed at obtaining evidence of and explaining the existing contractual relationships, at the provisions made in the contracts and at factors in the contracting environment, like organisational structures or the labour market.

[3] Although there is a connection between agency theory and the third party's agency function (the third party acts as an agent for the insured), it is important to notice that agency theory is not a theory about this agency function.

[4] Other approaches are, for instance, transaction-costs analysis and property-rights analysis.

Because of the methods used in the principal-agent literature and in the positive agency literature, Neelen (1993, p. 62) labelled the two types 'the analytical agency theory' and 'the empirical agency theory' respectively. Notwithstanding the differences, in both approaches the same problems of contracting within relationships with asymmetric information and conflict of interests are being addressed. Our approach to analysing relationships between third-party agents and GPs is mainly a positive one. While describing the way the financial and organisational arrangements between these parties should be structured, however, we will also use a more normative approach, although not mathematical.

After we have described the basic agency model, we will address some modifications of, or extensions to the traditional assumptions in subsection 3.2.3. Agency theory has been applied to, for instance, insurance arrangements, public bureaucracy, and employment.[5] Despite the different settings, the problems resulting from asymmetric information and conflicting interests are more or less comparable. These problems are discussed in subsection 3.2.4. Finally, we will present some examples of the omnipresent agency relationships in daily life (subsection 3.2.5).

3.2.2 Traditional assumptions of agency theory

Relationships between third parties and GPs are characterised by unequally distributed information. Each party has certain information the other party does not have. On the one hand, a third party may be at a disadvantage against a GP. For instance, a third party may not know the quality of the GP when entering into a contractual relationship. Further, although a third party may have some general information about an insured's health status, it is the GP who typically has more and better information about the clinical pictures of this insured and about the indicated diagnostic tests and treatments. On the other hand, a third party may have an advantage over a GP, for instance by having some performance figures of the GP's competitors that can be used in contract negotiations.

Another important characteristic of the relationships between third parties and GPs is that both parties may have different, and possibly conflicting, interests. A conflict of interests may give rise to problems in a relationship in which an ill-informed party wants a well-informed party to perform a certain action.

As to these problems, the relationships between third parties and GPs are not unique. On the contrary, relationships between two parties in which the one party has more or different information than the other party are omnipresent. Both parties may be involved in a transaction in which goods or services are being exchanged for some kind of compensation. Often, one party has been delegated to do some work for, or on behalf of

[5] Important contributors to the theory have been Spence and Zeckhauser (1971), Ross (1973), Jensen and Meckling (1976), Harris and Raviv (1978), and Shavell (1979). See, for instance, Mitnick (1980), Moe (1984) and Eisenhardt (1989) for other references and applications of agency theory.

the other party. The latter would like the first to act in his interests, but as both may have different information and different, possibly conflicting interests, it's obvious that problems can arise. A theory that focuses on relationships between two parties, that specifically addresses the problems of asymmetric information and conflicting interests, and that proposes strategies to deal with such problems (for instance, by suggesting optimal contract designs) is the theory of agency.

In agency theory, two parties can be distinguished. One party is supposed to be ill informed and is called the 'principal'; the other is supposed to be well informed and is called the 'agent'.[6] Generally, it is assumed that the agent may choose an action out of a set of possible actions and that this action, or its outcome, affects the principal's welfare as well as agent's. The kind of relationship they have may differ. Often, the 'agency relationship' is characterised as contractual, although there does not necessarily have to be a contract in the formal sense of the word.

The presence of an *asymmetry of information* between principal and agent is a crucial characteristic of an agency relationship. This asymmetry is due to the fact that in the first instance it is solely the agent who has information about his own actions, or about the circumstances upon which he bases these actions. In agency theory, this information is considered to be a commodity that can be obtained by paying a price (Eisenhardt 1989). Therefore, in general, the following assumption is made. The principal will not know which actions the agent has chosen to perform, how much effort the agent makes, or whether the agent has made the right decisions, unless he incurs costs to detect these actions or the information upon which they are based. Furthermore, there is no direct incentive for the agent to disclose this information to the principal (MacDonald 1984).

A second crucial characteristic of an agency relationship is the *conflict of interests*. Both the principal and agent have their own objectives and it is likely that these diverge. The principal would like the agent to act in his (the principal's) interests. However, besides affecting the principal's welfare, the actions of the agent also affect his own welfare. Both parties may value the agent's actions differently.

It is important to distinguish the concept of 'conflicting interests' from the concept of 'self-interest'. The first occurs within a relationship between two or more people, while this is not necessarily true for the second concept, as this is about human behaviour of an individual. In agency theory, it is assumed that both parties are driven by self-interest in an opportunistic or self-regarded manner.[7] Notice that conflicting goals are not much of a

[6] Notice that 'ill-informed party' and 'well-informed party' are relative concepts. Here, 'ill-informed' means less informed. The ill-informed principal may still be better informed than another party involved in, or affected by the transaction. Also, as Pratt and Zeckhauser (1985, p. 3) noted, an agent may know more about the tasks he has to perform, but a principal may know more about what he wants to be realised.

[7] Williamson (1985, pp. 47-50) distinguished different forms of self-interest seeking. In the weakest form, 'obedience', there is in fact no self-interest seeking. At the other extreme is 'opportunism'. Opportunistic behaviour includes deceiving people, giving them incomplete or false information on

problem if the agent is obedient. Likewise, opportunistic behaviour does not exclude harmonising goals.

What complicates the relationship is the presence of *outcome uncertainty*. Often, agency theory is applied to relationships in which the outcome of a particular process is uncertain and only partially the result of the actions performed by the agent. The outcome is also determined by factors over which the agent has no control, like technological developments, weather conditions or government regulation. The relative contribution of the agent to the final result is unknown. As a result, the principal may have problems drawing conclusions about the agent's effort from the outcome.

As noted, in agency theory agents are assumed to behave opportunistically. Next to this assumption, some other behavioural assumptions are being made. Agency theory is embedded in neo-institutional economics and, in general, it is assumed that both parties are rational, show utility-maximising behaviour and are risk averse (Neelen 1993). Often it is assumed that the agent is more averse to risks than the principal.[8]

3.2.3 Modifications of the traditional assumptions

The assumptions of the basic agency model about principal and agent are rather restrictive, which limits the applicability of the theory. In the course of time, some modifications of the theory, or extensions to it, have been proposed. A first one is the relaxation of the goal-conflict assumption. In some cases, the goals of principal and agent may converge, for instance, in case of employers and employees in a clan (Ouchi 1979) or in a small family business. [9]

Although the assumption of opportunistic behaviour of the agent is at the heart of agency theory, it is questionable whether agents always behave in that way. The

---

purpose. Somewhere in between is 'simple self-interest seeking' which is a more decent kind of behaviour.

Perrow (1986, p. 233) also discerned a 'neutral' kind of behaviour. Behaviour is considered neutral if one's action does not hurt or even helps another and entails no gain or loss to the actor. Also according to Perrow, 'other-regarding behaviour' helps the other and entails the actor a loss. Finally, in case of 'self-regarding behaviour' the other suffers a loss.

[8] Eisenhardt (1989, pp. 60-61) noted that the principal is assumed to be less risk avers than the agent because the first is able to diversify his risks. Such an assumption is, for instance, applicable to the relationship between employer and employee, or between insurer and insured.

[9] Employees (agents) who display a deep commitment to an organisation's objectives may improve their career prospects. However, the other way around, employees who face career prospects may display a deep commitment to these objectives. Hope of promotion may motivate an agent to respect the organisation's objectives (Sappington 1991). During the so-called socialisation process, objectives of an employee may be aligned with the objectives of the organisation. According to Ouchi (1979, p. 837), a clan is a unique organisation characterised by this socialisation process. He also distinguished between professions (a group of people who occupy different organisations with the same values) and cultures (the process of socialisation referring to all the citizens of a political unit).

relationship between patient and doctor or, again, the relationships within a small family business are examples in which this assumption might be relaxed.[10] Further, as Perrow (1986, p. 227) noted, it may be the principal instead of the agent who behaves opportunistically.

It is assumed that principal and agent are rational. Some, Eisenhardt (1989) for instance, assume that the rationality of the parties is bounded.

The programmability of the agent's tasks is another possible extension of the standard assumptions. Usually it is assumed that the agent's behaviour is difficult to observe and to evaluate. However, if a task is more programmed – task programmability is defined as 'the degree to which appropriate behaviour by the agent can be specified in advance' (Eisenhardt 1989, p. 62) – then observing and evaluating the agent will be less difficult.

In agency theory it is assumed that the outcome is the principal's most important source of information. The outcome may not be certain and it may be unclear how far it results from the agent's actions, but the outcome is observable or measurable. However, in practice the outcome may be difficult to observe or to measure. As Eisenhardt noted (1989, p. 62), it may take time before an outcome can be measured, or an outcome may be 'soft'. Related to this is the presence of several agents who all contribute to the outcome. Team effort makes it difficult for the principal to measure an individual agent's contribution to the outcome. If the agent is aware of the principal's problem, he may be tempted to shirk (Alchian and Demsetz 1972).

In case of more agents performing comparable tasks, the principal may use the relative performance of other agents as an indicator of a certain agent's efforts (Sappington 1991). The external factors over which the agents have no control should be the same for all agents then. Otherwise relative performance is unfair. Another advantage of more agents is that competition between agents may stimulate an individual agent to act in the principal's best interests.

Long-term or repeated relationships may lessen the information asymmetry in the course of time. The longer the principal knows the agent, the better he will know him and the circumstances under which the agent has to perform his actions, and the easier it will be to evaluate his behaviour. Also, the agent may be less tempted to cheat, as the risk of detection becomes larger. Moreover, the agent as well as the principal may have an interest in maintaining such relationships, because of investments that are made, experience or information that is acquired (to which there is a cost), et cetera.

Relaxation of the assumptions can have consequences for the design of the contract between principal and agent. If goals of both parties are similar, the agent does not behave opportunistically, the agent's tasks are programmable, there is competition between agents, or relationships are long-termed, then the contracts can be less complex.

[10] In the example of the small family business it is not clear whether the goal-conflict assumption or the opportunistic-behaviour assumption should be relaxed, though. The goals of the kids may resemble the goals of the parents, but they may also be obedient.

On the other hand, problems with observing or measuring the outcome, or the presence of several agents all contributing to it, may increase the need for more complex monitoring techniques.

3.2.4 Agency problems

The difficulties that may arise because of the combination of asymmetric information and conflicting interests are called 'agency problems'. Two types of agency problems can be discerned. The first problem is that of *hidden information*, *hidden knowledge* or, usually more specific, *adverse selection* (Arrow 1986). In this case the agent's actions themselves may be observable, but the information which he uses for deciding upon these actions is unknown. Therefore, the principal does not know whether, given the agent's knowledge, the agent has made the choice that benefits the principal most.

By adverse selection is meant that an applicant (the agent) usually is inclined to opt for an insurance policy in a way beneficial to himself but detrimental to the insurer (the principal). Because the applicant often has more information about his risk than the insurer, he may be able to choose a policy of which the price is lower than the expected claim costs. Adverse-selection problems only form a subset of the hidden-information problems. In the case of adverse selection the problem of asymmetric information occur before the contract is concluded, while in the other cases such problem also may occur during the contract period.

The second agency problem is called *hidden action* or, usually more specific, *moral hazard* (Arrow 1986). This hidden-action problem arises if a principal can not monitor the agent's actions, but only has some information about the outcome of the activity. The principal derives some utility from this outcome. Because the agent's actions and his effort increase the probability of the outcome the principal aims for, they are of value to the principal. It is assumed, however, that the agent's actions and efforts are a disutility to him (MacDonald 1984, Arrow 1986). Hence principal and agent may value the agent's effort differently.

In case of moral hazard, the agent's actions are not observable either. Often, these actions are induced by the presence of some kind of insurance. Because the agent is insured against a certain event or loss, he faces reduced incentives to take preventive actions or to minimise the damage at the moment the event or the loss occurs.[11]

Often, 'adverse selection' and 'moral hazard' are used in connection with insurance, although the terms are also used in other contexts.[12] Hereafter these terms will be used instead of 'hidden information' and 'hidden action'.

---

[11] Sometimes, it is the principal who is insured. In that case, moral hazard occurs in case the agent's actions are influenced by the principal's insurance. To this is returned in chapter 4.

[12] The terms 'moral hazard' and 'adverse selection' are, for instance, also used to describe problems within the relationship between employer and employee (see, for instance, Moe 1984).

3.2.5 Examples of agency relationships

Example A1: the relationship between insurer and applicant

Principal:	The insurer.
Agent:	The applicant.
Good/service:	Insurance against risk.
Desired actions:	Provision of information by the applicant about his risk.
Information asymmetry:	The applicant has more information about his risk than the insurer.
Conflict of interests:	The insurer would like to attract low-risk individuals, that is, individuals with expected claim costs lower than the premium charged for a certain policy, whereas the applicant would like to select a policy with a premium lower than his expected claim costs. The applicant has therefore no incentive to signal his risk voluntarily and truthfully.

Example A2: the relationship between applicant and insurer

Principal:	The applicant.
Agent:	The insurer.
Good/service:	Insurance against risk.
Desired actions:	Provision of information by the insurer about, among other things, his reliability, his policy on issues as service, payouts, investments and development of new products, his arrangements with intermediaries and repairers.
Information asymmetry:	The applicant has difficulties selecting an insurer, as it is the insurer who has more information about his reliability, policy, arrangements, et cetera.
Conflict of interests:	The applicant would like to select the insurer most fitted to his needs and wants, but the insurer has no incentive to reveal information in so far as this information is conflicting with his own interests.[13]

Before proceeding, the foregoing is illustrated with some examples of relationships in which the problems of asymmetric information and conflicting interests are present. It is shown that, from an agency perspective, relationships may be interpreted in several ways. The identification of principal (ill informed) and agent (well informed), their goals and the goods or services in question may be problematic. This is demonstrated by example A1 and example A2, which pertain to the adverse-selection problems within the

[13] For example, the insurer has no incentive to inform an applicant about the way he provides service if it leaves a lot to be desired (assuming that the insurer wants to attract the applicant).

relationship between an insurer and an applicant. Of course the desired actions as well as the interests of both principal and agent are hypothetical.

In example A1, it is the insurer who is ill informed. In the next example, the same parties are involved, but now the applicant is supposed to be ill informed. Further, the asymmetry relates to different information.

Example B: the relationship between firm and market researcher

Principal:	Firm.
Agent:	Market researcher.
Good/service:	Information about customer preferences.
Desired actions:	Collecting, analysing and providing information about what (potential) customers need or (want to) buy.
Information asymmetry:	The firm is unable to monitor the researcher and, therefore, does not know whether the researcher makes an effort to act in the firm's best interests. The firm has difficulties determining whether the final outcome is the best possible outcome.
Conflict of interests:	The firm has an interest in obtaining a high-quality research report at a low price, whereas the researcher may value income positively and effort negatively.[14]

Example C: the relationship between employer (or planner) and chauffeur

Principal:	Employer (or planner).
Agent:	Chauffeur.
Good/service:	Delivery of parcels.
Desired actions:	Delivering parcels as fast as possible.
Information asymmetry:	The employer or planner has no, or only little information about the way the chauffeur drives (speeding, losing his way), and also faces outcome uncertainty due to weather conditions and traffic jams.
Conflict of interests:	The chauffeur may have no incentive to act in the employer's or planner's best interests, that is speeding in order to deliver as many parcels as possible in the shortest possible time and thereby lowering costs[15]

[14] It is often assumed in the principal-agent literature that the agent derives a positive utility from income and a negative utility from effort. See, for instance, Shavell (1979), MacDonald (1984) and Arrow (1986).

[15] Whether the chauffeur has an incentive to deliver as many parcels as possible in a certain period of time, will, among other things, depend on the compensation in use.

Example B and example C concern problems of moral hazard. In example B, the principal is able to observe the outcome, which is the result of the agent's actions, but he is unable to evaluate the outcome. In example C, the principal's problems are increased due to the presence of outcome uncertainty.

3.3 Dealing with agency problems

3.3.1 Introduction

In agency relationships, the principal encounters problems of asymmetric information, conflicting interests and outcome uncertainty. As a consequence, the outcome – partially the result of the agent's actions – may deviate from the outcome the principal had in mind. The principal will have to motivate an agent to act on his behalf by drawing up a contract that stimulates the agent to choose an action that is in the principal's interests.

Subject of this section is the set of different strategies derived from agency theory that the principal can use to further the outcome he aims for. We classified these strategies into three groups. The first strategy is aimed at *selecting* the right agent. The second strategy is aimed at stimulating, informing and persuading, and directing the agent. The general term used here to cover these methods is *controlling*. The third primary strategy is *monitoring* the agent and involves observing or measuring his behaviour. Also subject of this section are the costs an agency relationship entails to both the principal and the agent.

3.3.2 Selecting the agent

In an attempt to minimise the agency problems, the principal may take some preventive measures. One possibility is agent selection, which serves two purposes.

The first purpose of selection is to reduce problems of diverging interests by seeking an agent who is committed to the principal's goals. Selection of a fitting agent takes place before the contract is concluded.

A second purpose of selection is stimulating the already selected agent to act in the interests of the principal. Simply the fact that the principal is in the position to select an alternative agent may serve as a stimulus to the agent to direct his behaviour at the principal. This 'threat of competition' takes place after the contract between principal and agent is concluded (Sappington 1991, p. 57).

There are different approaches to screen the potential agent. Firstly, the principal may try to distinguish the agents and classify them according to age, sex, training, occupation, or whatever relevant variable. Secondly, in case of agent characteristics that are more difficult to observe, the principal may stimulate a potential agent to reveal his true knowledge, skills, beliefs, expectations, et cetera. An indirect way of inducing an agent to disclose his hidden information is self-selection (Arrow 1986, p. 1187). By offering agents several contracts and letting them choose the one they prefer, the principal may be

able to differentiate the agents from each other. Finally, a principal may incorporate a probationary period.

In general, the selection problem will occur before a new contract is concluded. Another possibility is that the principal wants to deter the agent from terminating the relationship in order to prevent adverse selection problems. For instance, an insurer (the principal) may want to deter the insured (the agent) from terminating the insurance contract by raising switching costs or by colluding with other insurers (Schut 1995).[16]

Naturally, to be able to select an agent, a sufficient supply of agents is necessary. The problem, however, will often be the ability of the principal to identify among several agents the one who might serve the principal's interests best, and not so much the number of agents. Further, it is not necessarily the principal who selects the agent. The agent may also have been appointed to the principal (or the other way around) or the agent may have selected the principal. Finally, legislation may prohibit selective contracting or provide the agent with a monopoly position.

3.3.3 Controlling the agent

Another way to handle agency problems is trying to bring the objectives of the selected agent in line with those of the principal. The principal may try to influence the agent's behaviour and his efforts by using one or more methods of control. Although in agency theory (especially in the mathematical principal-agent literature) emphasis is placed on financial incentives and risk-sharing arrangements, other methods to control the agent are possible as well. These methods range from simply providing information hoping it will guide the agent, to forcing him by using (threats of) violence.

Distinctions between different controlling strategies have, for instance, been made by Lindblom and by Mitnick. Lindblom (1977) distinguished between control by exchange, control by authority and control by persuasion. Control by exchange involves inducing, in a market-like situation, an agent to perform an action by offering him a benefit in return. Control by authority involves unilateral co-ordination of an agent's actions within a hierarchical relationship or 'mutual adjustment' among different authorities.[17] Control by persuasion, lastly, involves transferring information to the agent in the hope that he

[16] Schut (1995, pp. 126-128) described several strategies a private health insurer may use to prevent adverse selection. These are rating risks (based on loss probabilities or loss experiences), selecting preferred risks (selective underwriting), stimulating risk signalling by agents (self-selection), raising switching costs, and collusion.

[17] Authority can also be exercised in absence of a hierarchical relationship. In case of mutual adjustment among authorities, different (equal) authorities give way to each other and exercise authority over each other. For instance, authority A may force authority B to co-operate by the threat of not co-operating with B's projects (see Lindblom 1977, pp. 29-31).

chooses the preferred actions or, somewhat stronger, persuading the agent to choose these actions.

Another classification was proposed by Mitnick (1980, p. 9) who made a distinction between regulation by incentives and regulation by directives, although he recognised that directives can also be viewed as negative incentives. Regulation by incentives involves 'changing the perception of the nature of the alternatives for action subject to choice; i.e. changing the relative attractiveness of alternatives.' Changing the relative attractiveness can be done by changing the alternatives or its characteristics, or by changing the agent so that he evaluates or perceives the alternative actions in a different way (Mitnick 1980, pp. 342-343). The agent, however, still has a choice which action to perform or how much effort to make. In case of regulation by directives, or rules for behaviour, the agent's choice is directed.

Neelen (1993, p. 71) combined the classifications of Lindblom and Mitnick. This resulted in three methods to control the agent: *control by incentives*, *control by persuasion or information*, and *control by directive or authority*. Neelen substituted 'control by incentives', analogous to Mitnick's 'regulation by incentives', for Lindblom's concept of 'control by exchange'. This is an arbitrarily decision since Mitnick's 'regulation by incentives' also contains methods comparable to Lindblom's concept of 'control by persuasion'. However, the resulting classification (control by incentives, control by persuasion or information, and control by directive or authority) is workable here. Firstly, Mitnick's 'regulation by incentives' is split into its two parts, i.e. changing the alternatives or its characteristics (control by incentives), or changing the agent (control by persuasion or information). Secondly, Neelen's classification contains the concept of 'incentives' which is a well-known concept in agency literature. Therefore, it is more workable here than Lindblom's classification. Hence hereafter Neelen is followed.

Control by incentives

Usually, the agent has to choose one action or a series of actions from a set of feasible actions. Control by incentives involves stimulating the agent to choose from this set those actions that benefit the principal most by making the preferred actions more attractive to the agent.

One can think of several ways to stimulate an agent. A first way to categorise possible incentives is by their nature. For instance, incentives can be financial or non-financial by nature. Non-financial incentives can be further subdivided into incentives that are non-financial but do have a monetary equivalent (like a larger lease car or presents) and incentives that do not (like compliments or admiration). In agency theory, incentives are usually stated in monetary terms. Financial incentives can be based upon behaviour (for instance, hourly wages as a proxy for behaviour), outcome (for instance, by letting the agent share in the yield) or a combination of both.

Often considered in agency theory – at least in the principal-agent literature – is letting the compensation of the agent depend upon the outcome by sharing the risks due to the external factors over which both principal and agent have no control. Does the outcome only result from the agent's actions, then there is no need for risk sharing and the

compensation can be based upon the outcome. In case of a random component, the risks will probably be shared. Whether risk sharing will be used in an agency relationship depends upon the principal's as well as the agent's aversion to risks. Suppose the principal is risk averse. Then a risk-neutral agent can be compensated by giving him the yield minus a fixed portion for the principal.[18] By so doing, the agent bears all the risk and has a strong incentive to make an effort. On the other hand, if the agent is averse to risks too, then sharing of these risks will be inevitable and the compensation will only to some extent depend upon the outcome (Shavell 1979). As a consequence of lessening the agent's risks, the incentive he faces to make an effort will be reduced too.

Apart from being financial or non-financial, incentives can also be positive (rewarding) or negative (penalising) by nature. Positive incentives entail, for instance, a bonus or a promotion in case the agent's behaviour or the results of his actions are satisfactory. Examples of negative incentives are a malus or termination of the contract in case his behaviour or his results are not satisfactory.

A second way to categorise possible incentives is by the fact whether they are made contingent upon a certain condition. Incentives may be conditional as well as unconditional on, for instance, the actions the agent performs or on his efforts. To illustrate this, consider the relationship between employer (the principal) and employee (the agent). In order to stimulate hard working, the employer may increase his employees' incomes. This increase is an incentive not conditional on the factual production level of an individual employee. Another possibility, however, is that the employer pays his employees on a piecework basis. The increase of an employee's income is made conditional on his efforts then.

To make the agent performing certain actions (or not performing them) on behalf of the principal, the latter has to draw up a contract that stimulates desired behaviour: the 'incentive compatibility constraint'. The stronger the incentives, the more the agent will be forced to act in the principal's interests. However, the principal often is limited in his choice of contract design. If the agent considers the contract as unattractive, he may not want to accept it. Moreover, in case of more than one principal there may be competition for the agent. If there is, the agent may have several options to choose from. Hence the contract offered by a certain principal must be at least as attractive as those offered by other principals. This limitation is called the 'participation constraint' (Arrow 1986, Ryan 1994).

Control by persuasion or information

Whereas incentives stimulate the agent to perform the preferred actions, control by persuasion or information involves informing the agent so that he values the preferred actions more positively and persuading him to perform these actions. The principal may

[18] This fixed portion can be seen as a 'franchise fee' (Sappington 1991, p. 47).

gather data and inform the agent directly, use advertising, point at societal norms, et cetera. In principle, the rewards and penalties remain the same though.

Control by directive or authority
Control by directive or authority differs from the other two controlling strategies in that the agent's choice is restricted. The actions the agent has to perform are defined. As do control by incentives and control by persuasion or information, control by directive or authority of course leaves room for undesirable actions performed by the agent, but he will have to bear the costs of sanctions then. In case of strong incentives there is a resemblance to control by directive or authority for in fact the agent's choice is restricted too.

In agency theory authority and hierarchy is sometimes left out of consideration and both parties within the relationship are considered to be equal. That is, there is no hierarchical relationship and contracting takes place on the market instead of, for example, within organisations. For some relationships, for instance the relationship between an insurer and an insured, this will not present problems. For other relationships, however, these concepts cannot be ruled out. An example of such a relationship is the one between employer and employee. Therefore, the absence of authority and hierarchy in some relationships has been seriously questioned by organisational theorists (see, for instance, Perrow 1986) but has also been recognised by agency theorists (see, for instance, MacDonald 1984). Pratt and Zeckhauser (1985) noted that, as it is assumed that monitoring the agent is costly, organisations might fulfil several control functions that can not be achieved in market-type relationships as well. According to Pratt and Zeckhauser (1985, p. 32) 'forming an organisation only internalises agency relationships. It does *not eliminate* problems of co-ordination, incentives, and so on. Presumably, though, it should substantially *lessen* them, for now we are all working on the same team and can be rewarded accordingly' (italics added, A.V.).

3.3.4 Monitoring the agent

There are different ways to handle the problems of imperfect information that occur after the contract is concluded. One way is to monitor the agent in order to reduce the information gap. Besides trying to acquire information about the agent's actions, the principal may also try to trace the information the agent uses for deciding upon these actions.

A different approach might be to try to acquire information about the external factors which influence the outcome but over which the agent has no control. If the principal succeeds in this, he may be more able to draw conclusions about the agent's actions from the outcome. Hereafter, the term monitoring will be used solely to denote observing or measuring the agent's behaviour. Acquiring information about the external factors over

which the agent has no control will be denoted as monitoring, observing or measuring the external factors.[19]

Monitoring the agent serves several purposes. Obviously, it may reduce the principal's informational disadvantage, but it can also be viewed as an incentive. The agent, knowing his behaviour is being observed or measured, may be stimulated to act in the principal's best interests if his compensation is contingent upon this behaviour (see subsection 3.3.3). Further, investment in observation and measuring instruments is one of the preventive measures the principal may take before the contract is concluded in an attempt to minimise agency problems. The presence of such devices may discourage a potential agent who has the intention to cheat to enter into a relationship with the principal. It thus can be seen as an additional way to select an agent.

3.3.5 Agency costs

Because the transaction costs of selection, controlling and monitoring are positive, the principal has to incur costs to accomplish desired behaviour. But, as will be explained, also the agent may have to incur some expenses.

Different types of costs can be distinguished. Mitnick (1980, p. 150) distinguished between specification costs and policing costs. *Specification costs* are the costs the principal has to incur to identify the actions and efforts that serve his interests the best. The agent may also have to incur such costs '(...) in ascertaining and acting for the principal's preferences.' The principal has to pay *policing costs* to monitor the agent and to enforce compliance.

Jensen and Meckling (1976, p. 308) distinguished between three types of costs, of which the sum is labelled 'agency costs'. Firstly, the principal has to make *monitoring and controlling costs* to reduce the information asymmetry by observing or measuring the agent's behaviour and to control the agent.[20] Secondly, the agent may take measures to guarantee that he does not harm the principal. Also, the agent may ensure that the principal is compensated whenever he does act contrary to the principal's best interests. The expenses the agent has to incur are called *bonding costs*.[21] If in spite of the incurred monitoring and bonding costs the outcome indeed deviates in a negative manner from the

[19] Elsewhere, however, the term monitoring is also used in a broader sense: Jensen and Meckling (1976, p. 308) understood by monitoring not only observing or measuring but also controlling the agent, for instance by using incentive schemes. For reasons of uniformity, the term monitoring and controlling costs will replace the term monitoring costs as used by Jensen and Meckling (see subsection 3.3.5).

[20] Jensen and Meckling (1976, p. 308) labelled these costs monitoring costs (see the previous footnote). As examples of monitoring methods they mentioned auditing, formal control systems, budget restrictions, and incentive compensation systems (1976, p. 323).

[21] Examples of bonding methods are a contractual guarantee that the financial accounts will be audited by a public account, bonding against malfeasance of the agent, and limitation of the agent's decision making power (Jensen and Meckling 1976, p. 325).

outcome the principal aimed at, then is spoken of the *residual loss*, which is the dollar equivalent of the principal's welfare reduction.

Dealing with agency problems implies searching for structures that can reduce the residual loss. However, reducing this loss requires incurring monitoring and controlling costs as well as bonding costs. Total agency costs, then, have to be minimised by finding a level of equilibrium between monitoring and controlling costs plus bonding costs on the one hand and the residual loss on the other hand.

3.4 Using agency theory

In this study, agency theory serves as the theoretical framework by which the relationships between third parties and GPs are analysed. Although the theory seems well suited to serve as theoretical framework here, the choice of agency theory entails some limitations. Firstly, the theory of agency itself has some limitations. Secondly, the agency theory may not be fully applicable to the relationships between third parties and GPs. The first limitations are subject of this section; the latter are discussed in the next chapter.

As with all theories, the theory chosen here is a simplification of the surrounding world. An example is the isolated environment in which agency relationships, especially in the principal-agent literature, often are assumed to exist. It is practically impossible to deal with all the variables that affect the decisions of both principals and agents. But even though models used in agency theory tend to be rather simplistic, sometimes the proposed solutions (complicated financial contracts) are often more complex than observed in practice. Arrow (1986, pp. 1193-1194) mentioned three reasons why such complicated fee functions are not found in reality. Firstly, both drawing up the arrangements included in complex contracts and living up to them is costly. Thus the more complex the contract, the higher the transaction costs. Another reason is that in reality all kinds of subtle monitoring means are present that make it less necessary to devise complex incentive mechanisms. Thirdly, in reality several incentives replace the complicated financial incentive structures proposed by agency theory; non-financial stimuli may also influence the agent's behaviour. Examples of these are compliments, appreciation, pride, social and professional norms, peer review, or dismissal.

Another questionable point is the assumption that the agent behaves opportunistically. Although it is important to recognise that some agents may behave opportunistically, it is more interesting to investigate under which conditions agents behave in a self-regarding way and when they behave in a neutral or other-regarding way (Perrow 1986, p. 232).[22] Perrow (1986, p. 234) also noted that '(...) there is no innate tendency to either self- or

[22] Opportunistic behaviour is also self-regarding behaviour, but self-regarding behaviour does not have to be opportunistic. The latter includes deceit while self-regarding behaviour may be more decent (see subsection 3.2.2).

other-regarding behaviour in people; either can be evoked depending on the (organisational or societal, A.V.) structure.'[23] He continued by stating that agency theory '(...) gives scant attention to the co-operative aspects of social life (...)' (1986, p. 235).

Another behavioural assumption concerns the maximising behaviour of both principal and agent. As in the neo-classical theory of the firm, in the neo-institutional agency theory actors are viewed as utility maximising. However, different levels of rationality can be distinguished, ranging from maximising behaviour, via bounded rationality, to organic rationality (Williamson 1985, pp. 44-47). There is no unanimity about this human-behaviour assumption.

Despite these limitations, agency theory has been chosen as the framework. The main reason is that the central problems described by the theory seem to resemble the problems that occur in the different relationships between third parties and GPs very well (see also section 3.1). In the existing relationships, an agent (the GP) provides goods or services and is compensated by a principal (often a third party). Furthermore, the relationships seem to be characterised by asymmetric information and conflicting interests. Also, the outcome of health care is often uncertain due to external factors over which the provider has no control.

3.5 Summary

In this chapter, we discussed the (choice of a) theoretical background for our study of the various relationships between third-parties and GPs and of the various strategies the third party may use. There are several reasons for choosing an agency perspective instead of, for instance, focusing on the allocation of decision rights. One reason is that agency theory seems to correspond well with the situation central in this thesis concerning the contractual relationships between two parties (that is, third-party agents and GPs) in which the one party (the principal) enters into a relationship with a second party (the agent) in the expectation that the agent's actions are beneficial to the principal. Further, the aspect of the use of (financial) incentives corresponds well with the subject of this thesis, namely financial-risk sharing.

Although a decision rights approach could shed an interesting light on the different kind of organisational arrangements between third parties and GPs, questions of integration and ownership of assets are beyond the scope of this thesis. Furthermore, the use and the ownership of assets are less important facts in general practice than they are for instance in medical specialist care, and there are hardly relation-specific assets.

[23] Perrow (1986, p. 233) mentioned several organisational conditions which favour self-regarding behaviour, for instance, minimisation of continuing interactions (no long-term relationships), encouragement of storage of rewards and surpluses by individuals and of measurement of individual effort, minimisation of interdependent effort through choice of equipment and work-flow design.

Regarding the question of authority, we assume that the professional or the physician's individual autonomy makes it difficult for a third party to interfere in the patient-physician relationship in a way different from the ways in a market relationship. Hence we assume that the professional relationship with the GP makes that the agency problems within an employment situation do not significantly differ from the agency problems within a market relationship. A final consideration is that GPs are often self-employed.

We have given an overview of agency theory in order to provide an answer to the second research question:

What are the main characteristics of agency theory?

In agency theory, a principal commissions an agent to perform some actions for, or on behalf of him. The relationship between both parties is characterised by asymmetric information and conflicting interests. The agent has more information about his intentions, his actions and his efforts or about the circumstances upon which he bases his actions and may have different goals. As a result, problems of adverse selection and moral hazard may arise. Moreover, the principal may have problems drawing conclusions about the agent's efforts by observing the outcome. This is due to the fact that in many agency relationships the outcome is uncertain because of external factors that influence the outcome but over which the agent has no control. In order to achieve the outcome he aims for, the principal has to design a contract that motivates the agent properly.

Within the theory of agency, two approaches can be distinguished: the normative and non-empirical principal-agent approach, which is highly mathematical, and the empirical non-mathematical approach of the 'positive theory of agency'. In both approaches authors concentrate on contracting problems between two parties in a relationship with unevenly distributed information and conflicting interests. Also, in both approaches assumptions of conflicting goals, human behaviour (opportunism, rationality and risk aversion), and information (a commodity, and asymmetrically distributed between principal and agent) are being made. While describing mechanisms or techniques third parties may and actually do use within their relationships with GPs, our approach can be characterised as a positive one. While discussing how systems of financial-risk sharing should be structured, our approach can be characterised as normative, although not mathematical.

Some modifications of the traditional assumptions of the basic model of agency have been proposed. These relate to the assumptions of goal conflict and opportunistic behaviour, the rationality of principal and agent, the programmability of the agent's tasks, the measurability of the outcome, the number of agents, and the duration of the relationship.

There are different strategies to handle the problems of agency relationships. Firstly, the principal may select an agent who is expected to behave in the principal's best interests. Then, the principal may use different strategies: controlling the agent and monitoring the agent. Different controlling strategies are control by incentives, control by persuasion or information, and control by directive or authority. By monitoring the agent,

the principal may reduce the information gap and may also stimulate the agent to act in the principal's interests.

To accomplish desired behaviour, the principal has to incur monitoring and controlling costs. The agent may have to incur bonding costs. Nevertheless, there still may be a residual loss due to the divergence between the realised outcome and the outcome the principal had in mind. Total agency costs are minimised by searching for a level of equilibrium between the costs of selecting, monitoring and controlling the agent on the one hand, and the residual loss on the other hand.

In spite of some limitations the theory seems well suited to serve as the theoretical framework here.

4 AGENCY AND HEALTH CARE

4.1 Introduction

In the previous chapter it has been argued that agency theory is applicable to relationships between two parties (a principal and an agent) that are characterised by asymmetrically distributed information and conflicts of interests. Although agency theory is not specifically aimed at the various relationships in the health care sector, the agency characteristics, which may be found within such relationships, make its application to that sector reasonable. On the other hand it may be argued that the health care sector has some special characteristics that may limit the applicability of the theory.

Two research questions are addressed in this chapter. In the first place is dealt with the fourth question as formulated in chapter 1:

Is agency theory applicable to relationships in the health care sector in general, and to the relationships between third-party agents and general practitioners in particular?

To that end, some specific relationships in the health care sector are analysed using agency theory. We start with what might be viewed as the central relationship in health care: the patient-physician relationship. In the first instance (section 4.2) this relationship is considered in isolation from other relationships, that is, as if no third party is involved. As argued in chapter 2, however, the health care sector is characterised by the presence of a third party performing the insurance function and the agency function. Hence in the second instance the third party is introduced. This introduction results in new agency relationships. The relationships dealt with here are those between patient (i.e. insured) and third party (section 4.3) and between third party and GP (section 4.4). The patient-physician relationship is affected by the presence of a third party and, therefore, is reconsidered in section 4.5. In subsection 4.6 we use the several concepts derived from agency theory to construct a theoretical framework of which the relationship between third party and GPs is a part, and that provides the rationale for the third party's strategies to influence the GP.

4.2 Patient and general practitioner

4.2.1 Introduction

The relationship between patient and physician may be considered as the main relationship in health care. It is true other relationships in health care are also important, but in general they only facilitate the patient-physician relationship. The question to be answered here is whether the theory of agency is applicable to the relationship between a

patient and a GP. In practice, the amount of information both parties possess as well as their preferences and attitudes will differ per patient and per physician. But even the relationship between a specific patient and physician may differ per visit and depend on the nature of the patient's complaints. For instance, the relationship a chronically-ill patient may have with a physician will change once an acute disease has developed.

In subsection 4.2.2, agency theory is applied to the relationship between patient and (primary care) physician. In subsection 4.2.3, some modifications of the standard assumptions are discussed. The concept of perfect agency is discussed in subsection 4.2.4. Finally, in subsection 4.2.5 we analyse the agency problems that may occur within the relationship.

4.2.2 Applying the theory of principal and agent

Agency theory seems applicable to the patient-physician relationship. Indeed, several authors described the relationship using agency theory (see, for instance, Evans 1984, Dranove and White 1987, Blomqvist 1991, Mooney and Ryan 1993, Ryan 1994, Scott 1996, Smith et al. 1997) or mentioned it as an obvious instance of an agency relationship (see Moe 1984, Arrow 1986, Spremann 1989). According to Arrow (1986, p. 1193), the patient-physician relationship is even an almost perfect example of the principal-agent problem.

In agency terms, the relationship is characterised by *asymmetric information*. The GP is supposed to be the well-informed agent, the patient the ill-informed principal. The patient may have information about what he wants to be realised, but usually it is the physician, as a trained and skilled professional, who is much better informed about clinical pictures, possible treatments and the effects or side-effects of specific treatments. Does the physician already have a lead at the beginning of the relationship, the difference in information may increase as the relationship develops and the physician obtains more and more information about the health status of a particular patient. But then, the patient may also acquire (part of) this information and thereby keep up with the physician. Whether the patient will be able to reduce the asymmetry of information is open to question.

The extent of the difference in information between patient and physician may differ per relationship. The patient may have some experience with a certain disease and may be almost as well informed as the physician, which is especially plausible in case of a patient with a chronic disease. If the difference in information is minor or even absent, continuation of such a relationship may still make sense if the patient depends on the physician, for instance for a prescription or a referral. In absence of an asymmetry of information, however, the relationship doesn't meet the characteristics of an agency relationship anymore.

Further, not only the patient but also the GP may not always be well-informed. Due to a lack of knowledge or experience he may not know which actions to perform or whether his actions are effective, considering the patient's symptoms. A shortage of knowledge may be a limitation of a particular physician, but it may also reflect the absence of certain knowledge or consensus within the medical profession. As Pauly (1978, p. 17) remarked, 'no one knows whether board certification, tonsillectomies, or some lab tests will im-

prove health outcomes or not.' Patients are often not able to evaluate the quality of care, that is, the quality of treatments, diagnostics, prescriptions or referrals, but sometimes physicians are neither. According to Evans (1984, p. 89), there is, 'particularly for diagnostic and monitoring activities, a broad zone of uncertainty in which optimal treatment and the limits of efficacy have not been scientifically established.' Although in case of a relatively ill-informed physician the extent of the information asymmetry may be limited, the patient may still think such an asymmetry exists. Only the idea that the physician is better informed may cause agency problems.

The second crucial characteristic of an agency relationship, the presence of a *conflict of interests*, may also be found here. The patient will expect the physician-as-agent to act in his best interests by providing goods or services properly. In case of a GP, examples of these goods and services are information, advice, prescriptions, referrals, medical services provided by the physician in question, and co-ordination of care provided by other providers. In return for this provision, the physician will receive a financial compensation. However, the relationship may be characterised by conflicting interests. Once the patient has made a price-quality trade-off, he may want the physician's unconditional attention, devotion and effort, whereas the physician's behaviour is restricted by his own objectives. Although the physician will probably take the patient's interests into account, acting accordingly may reduce his own welfare. The physician's welfare may, for instance, be a function of his income, his leisure, his workload, and his intellectual satisfaction (Scott 1997). Whether and how the physician's income is influenced by his actions depends on the compensation scheme. Then the question arises how income affects the physician's welfare. Probably, it will have a positive effect but with a diminishing marginal utility.

Common ways of compensation are fee for service, capitation and salary. Each action by the physician may increase his income in a fee-for-service system, whereas it may relatively decrease his income in a capitation system. In a salary system, the income will remain unaffected. By using a certain compensation system, the patient can thus stimulate the physician to act in the desired manner.[1] The conflict will depend on the compensation scheme in use. For instance, in a fee-for-service system the interests of patient and physician are similar as long as the marginal benefits of care exceed the marginal costs for both physician and patient. Once the marginal benefits for the patient become lower than the marginal costs (i.e. fee plus time costs), a conflict may arise. From this point on, the physician may want to provide more care than is best for the patient if the fee is higher than the marginal costs of providing care. In a capitation system the incentives are different. Here a conflict may arise as long as the marginal health effects are positive and will disappear once the marginal effects diminish. In case of zero or negative effects, neither the patient nor the physician has an incentive to demand or to provide care.

Another major concept of agency theory, the presence of *outcome uncertainty*, seems to apply to the present relationship too. The outcome is uncertain and will only partially

[1] Notice that in this section the patient-physician relationship is considered in isolation from other relationships, that is, in absence of a third party. It is not very likely then that the patient would opt for a salary or a capitation scheme.

be the result of the GP's actions. It will result from the course of medical treatment as well as the natural course of the disease or other health influencing factors. As a result, the patient will have problems assessing the appropriateness of the physician's actions and the quality of the physician as such.

Because of the differences in information and interests, the patient can not assume that the GP will always act in his interests. Therefore, the patient has to devise a contract that will stimulate the physician to act in a way that is beneficial to the patient. As the physician will not accept each contract and, moreover, will probably have different contracts to choose from, the contract offered by a particular patient will have to be at least as attractive as those offered by other patients. The options the patient has then range from what he finds minimally acceptable to himself – this will be the result of a price-quality trade-off – to what is minimally acceptable to the physician. The latter limit is called the 'participation constraint' (Arrow 1986, p. 1189). Eventually, the patient can choose the option of 'voting with his feet' (exit) if the costs of staying with the same physician outweigh the transaction costs of switching (Tai-Seale 2004). He will have to find another physician then who is prepared to accept his contract terms.

Usually, the physician is depicted as the agent. To illustrate that the role of both principal and agent might also be reversed, another way to look at the relationship is described here. By reversing the roles of both parties, the physician now becomes the ill-informed principal, the patient the well-informed agent. There seems to be an asymmetry of information in the sense that it is the patient who has essential information about his medical history, current health status, life style, true intentions for visiting a physician, et cetera. The ill-informed physician would like to obtain this information, as it may be crucial to the decisions about diagnosis and treatments. Also, the physician would like to have the patient acting in his interests by complying with his prescriptions. As will be shown hereafter, these actions are in the interests of the physician because of his reputation, for financial reasons, et cetera. The patient-as-agent, however, may have own interests. Especially in the pre-contractual phase he may want to withhold information that might influence the terms of the contract. Further, he may have his own reasons for paying a visit, which don't necessarily have to be medical by nature. He may, for instance, want attention or he may want to legitimise his absence from his work. Further, the patient may have his reasons for holding back information, or for distorting it. The information may, for instance, be intimate by nature and the patient might feel ashamed about it. Likewise, provision of certain information may reveal the patient's lifestyle, which may be a major cause of his symptoms. Guilt then may restrain the patient from telling the truth. As to the compliance with the physician's prescriptions, it is well known that compliance is far from optimal.

Both parties may have both roles; it is not either the patient or the physician who may fulfil the role of principal. Moreover, it could be argued that they are principal and agent at the same time. The GP is agent in his relationship with the patient in which he possesses more information about clinical pictures, treatments et cetera, and in which he is supposed to apply his knowledge and experience in a way which increases the patient's

welfare most. The physician is principal in his relationship with the patient in which the latter is better informed about the possible causes of his symptoms, the nature and the extent of this symptoms, his true intentions et cetera. In both cases, the patient and the physician have their own interests, which will probably diverge.

4.2.3 Modification of standard assumptions

Although there is a clear resemblance between the patient-physician relationship and the standard agency relationship, the first diverges from the latter in several respects. Therefore, some assumptions of the agency theory are relaxed here.

Perhaps the most noticeable difference between the patient-physician relationship and the standard agency relationship concerns the *conflict of interests*, which may be less severe in the first relationship than in the latter. In agency theory the agent is supposed to act in his own interests, which may be problematic unless his interests are, for instance by coincidence or due to incentive mechanisms, similar to those of the principal. However, Evans (1984, p. 79) argued that a professional relationship, like the patient-physician relationship, should not be confused with the standard principal-agent relationship. 'In the principal-agent problem (...) the *objectives* of principal and agent are strictly separated; each intends to serve only her own interests. What distinguishes the professional agency relation is that the professional includes part at least of the patient's/client's interests in her own objectives.' Professional norms and values taught him during his training and social control by peers, patients, et cetera, will stimulate the physician to direct his actions at the patient. Nevertheless, it is likely that there is a large variation in this altruism across the population of physicians (Jack 2005). Also important here is the notion of trust (Gray 1997). The patient will have to trust in the physician's capabilities and trust that he applies his skills in a way that is beneficial to the patient, even if this is incompatible with the physician's own interests.[2]

Another reason why the conflict of interests may be smaller is the nature of the relationship between a patient and his GP. Especially the relationship between a patient and a GP may often be characterised as a lengthy or a repeated contact. This is in contrast to the relationship with a specialist, which is often relatively short-termed. A long-term relationship may prevent the physician from pursuing solely his own interests, because poor performance may prompt the patient to terminate the relationship or not to renew it (Dranove and White 1987). Poor performance may also ruin the physician's reputation. Further, Jelovac (2001) argued that if repeated contracts are a possibility and depending on the type of contract between third party and physician, the physician may be induced to recommend the most adequate treatment in order to avoid additional treatments and their associated costs in the future. The same will hold for the present relationship between patient and physician.

[2] The patient as the ill-informed principal has to trust in the physician's knowledge and experience, and thus in fact has to trust in the *presence* of an information asymmetry. As to the conflicting interests, he has to trust in the *absence* of a conflict.

A next remark about the differences between both types of relationships pertains to the *principal's goals*. In agency theory it is assumed that the principal knows what he wants to be realised. In the patient-physician relationship, however, it is questionable whether the patient-as-principal indeed knows what the final result of the physician's actions should be. The patient may have some general ideas, like 'control the disease' or 'restore health', but often it will be the physician who will determine which goals are attainable, and which are not.[3] Moreover, often the goals can only be set after the patient has been diagnosed, that is, after the relationship has been established. This may have consequences for the strategies the patient may use to select and to control the physician. For instance, it is hard to select an agent if it is not clear which knowledge and experience are required at the moment the selection takes place. Also, without the diagnostic information it will be difficult to make the actions that are mostly preferred by the patient attractive to the physician.

Two final remarks pertain to the *outcome* of the physician's actions. Firstly, the outcome is non-financial by nature. Instead of in financial terms, it will probably be stated in terms of health status or well-being. Secondly, as noted above, the outcome is uncertain due to the natural course of the disease and external factors over which the physician has no control. But what distinguishes the present relationship is that the patient also influences the outcome. Although it is not uncommon that a principal has some influence on the outcome, the extent of the patient's influence seems to be significant.[4] Not only the provision of information by the patient determines to a large extent the success of the physician's actions; the patient's compliance with the physician's recommendations or prescriptions is another major outcome determinant.

4.2.4 Perfect agency

Having recognised that the GP as the patient's agent may possess superior information and have different interests, the question arises when he acts as the patient's 'perfect' agent. In the discussion about a perfect-agency relationship, two matters are put forward. Firstly, there is the question of who should be the decision-maker. The physician may convey the information to the patient and let the patient make the decisions, or he may make the decisions himself and perform those actions which he thinks are in the patient's best interests. The second question pertains to what should be enhanced: the patient's health or his well-being.

There are several views on the physician's agency role. Evans (1981) viewed the relationship between a patient and a GP as one in which the physician has, to some extent, integrated forward into the role of the patient. Perfect agency requires complete forward

[3] For the patient it may be hard to judge the need for and the quality of the physician's goods and services. Often, it will be the physician who sets the goals and who decides in which way these goals may be achieved. Hence the physician's goods and services have the characteristics of credence goods.

[4] That in a standard agency relationship the principal may have some influence on the outcome can be illustrated by example C in section 3.2.5. The outcome is the result of the chauffeur's actions and external factors, like weather conditions or other traffic, but it also depends on the employer for it is the latter who decides upon matters like the purchase of reliable cars or communications equipment.

integration, whereby the individual patient and the physician form an informed pair. This informed pair should use the physician's information and the patient's objectives and constraints to take decisions that are solely in the patient's best interests. But despite this pairing, elsewhere Evans (1984, p. 75) viewed the physician as the agent '(...) trying to choose what the patient would have chosen, had she been as well-informed as the professional.' Thus, it seems to be especially the physician who is making the decisions on the patient's behalf. What is in the patient's best interests is not entirely clear. Evans spoke of maximising the patient's well-being, but he recognised that it may be beyond the physician's ability to do so. The physician's primary goal will be to deliver or to arrange care in order to improve the patient's health. Evans distinguished three types of agents. The agent in the standard agency relationship will always try to increase his own welfare and behave opportunistically. The professional-as-perfect-agent will not dupe the patient but direct his behaviour at him. In the middle of these extremes is the professional-as-imperfect-agent who will sometimes act in the patient's best interests and sometimes not.

Culyer (1989, p. 37) expressed a somewhat different idea on perfect agency. His view was similar in that he considered the physician as an agent 'ideally choosing in the way the individual would, had he or she been possessed of the same informational advantages as the professional'. At least part of the decision-making about health-enhancing actions has been delegated to the physician here. The major difference is that according to Culyer it is health that generates utility instead of well-being.

A different view is that '(...) the doctor's role is to give the patient all the information the patient needs in order to enable the patient to make a decision, and the doctor should then implement that decision once the patient has made it' (Williams 1988, p. 176).[5] According to Williams, there is no delegation of decision-making from the patient to the physician in case of perfect agency; it is the patient who decides.

Perfect agency requires similar interests, or at least a physician who acts according to the patient's aims. In case of perfect agency, then, it does not make a lot of difference whether the physician or the patient himself makes a decision on the latter's behalf.[6] The decision will probably be the same. As the aims are the patient's, the physician-as-perfect-agent should behave altruistically, but not paternalistically. Does, despite the equal information, the patient's opinion about what the proper decision is deviate from what the physician would have decided had he been the patient, then the physician should not 'overrule' this opinion. What is 'perfect' depends on one's viewpoint; in agency theory it is the principal's, i.e. the patient's, and not the agent's.

In short, there does not seem to be agreement on what perfect agency is about. The GP as perfect agent may be viewed as facilitating the patient's decisions, or as the one making the decisions on behalf of the patient. It is also unclear whether the physician is sup-

[5] Williams continued by arguing that interchanging the terms 'patient' and 'doctor' seems more in accordance with reality. The result becomes then 'the *patient's* role is to give the *doctor* all the information the *doctor* needs in order to enable the *doctor* to make a decision, and the *patient* should then implement that decision once the *doctor* has made it' (Williams 1988, p. 176).

[6] It will make some difference whether it is the physician or the patient who is the decision-maker if the patient derives some utility from the decision-making as such.

posed to act in the patient's best interests by enhancing the patient's health or his well-being.

A distinction can be made between a theoretical view on perfect agency on the one hand and a more practically oriented view on the other hand. In the theoretical view, patient and physician indeed form a pair, like Evans (1981) described it. The patient informs the physician of his aims. Further, he provides all the information he possesses about his medical history, his current symptoms, his life style, and other possibly relevant information. The perfectly-informed physician – that is, perfectly informed about his field – uses this information while deciding upon the most desirable actions, after which he discusses these actions with the patient. Both parties become perfectly informed; the patient acquires the physician's knowledge and vice versa. In this interactive process, both patient and physician may want to adjust their aims and actions to the acquired information. For instance, the patient may adjust his aims due to the acquired diagnostic information, while the physician may adjust his plans to the patient's aims. Such a way of exchanging and discussing information and shared decision-making resembles the legal doctrine of 'informed consent' (President's Commission[7] 1982).[8]

A perfect-agency relationship as described is practically unfeasible, as the perfect-information condition will not be met. Physician and patient are not perfectly informed. In case the physician does not inform the patient, the latter will have problems becoming informed about diagnosis, intervention plan, et cetera (Starfield 1992). The patient will not be able to adjust his aims then. In the more practically oriented view on perfect agency, the physician will still do his utmost to act in the patient's best interests, given the limited knowledge in his field or given constraints over which the physician has no control (like legal constraints or the patient's limited financial means). A physician who purposely withholds information or is negligent in acquiring information is no perfect agent; not in the theoretical view, but not in a more practical view either. There are several reasons a physician may not want to inform a patient (Ryan 1994). The physician may have the idea that a patient does not want certain information, or that he will not understand it. Time constraints may also restrain the physician from informing the patient. Further, the physician may not want to upset the patient, or he may be afraid that the patient will reject the proposed actions. From an agency standpoint, however, these reasons imply an imperfect agency relationship.

The perfect-information condition will also not be met if the physician lacks knowledge or experience in general. Further, there may be a communication problem resulting in a lack of information about the patient in question. Firstly, instead of anamnesis the physician may use other ways to acquire information and by so doing miss some details. Secondly, he may not give the patient enough time to provide the information. Thirdly, he may miss some information by directing the conversation too much (Starfield 1992).

[7] President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, shortened to President's Commission.

[8] As communication is essential, perfect agency cannot be accomplished when the patient is comatose or has been anaesthetised.

A second reason why a perfect-agency relationship as described is not feasible is that the similar-interests condition will not be met. The interests will diverge as the physician has his own interests, which cannot be ruled out and which may result in a conflict of interests. For instance, the way the physician is compensated may contribute to this conflict as it may induce him to provide more or less services than what could be considered as optimal from the patient's point of view. If the physician makes a compromise between the patient's interests on the one hand and his own interests on the other hand or if he deliberately withholds information, then he acts as an imperfect agent.

An important assumption made in the foregoing is that the patient has the capability to make health care decisions. This requires that he possesses a set of values and goals, is able to communicate and to understand the information, and is able to reason and deliberate about his choices (President's Commission 1982, p. 57). For some patients, these conditions will not be met. Is a patient incapable of making health care decisions, then alternatives, like appointing a guardian, have to be sought. As this is beyond the scope of this thesis, such patients are not considered here.

Another reason for departing from the usual way of decision making is if the patient has stated explicitly – by means of a waiver – that he does not want to be informed. Then, withholding the patient information is in his interests and thus valid.

Other examples of situations in which informed consent is not required are medical actions which are directed or authorised by the law (legal requirements), emergencies, and the so-called therapeutic privilege in case informing the patient would be detrimental to the patient (President's Commission 1982, p. 93).

Summarising, in case of perfect agency it would not matter whether the patient or the GP would be the decision-maker. Both would become perfectly informed and both would act in the patient's best interests. In practice, however, for several reasons the perfect information condition will not be met and there will be a problem of conflicting interests. Hence, the patient will have to draw up a contract by which the physician is motivated as much as possible to act in a way advantageous to the patient. Whether such a contract should stimulate the physician to enhance the patient's health or his well-being is open to question. There may be some truth in Evans's consideration that maximising the patient's well-being may be beyond the physician's ability. On the other hand, it may be argued that especially GPs are confronted with problems that are related to well-being rather than health. This is beyond the scope of this thesis, however.

4.2.5 Agency problems

Due to the different information and interests, the two agency problems of adverse selection and moral hazard may occur. The patient may be confronted with adverse selection if he has problems assessing the physician's knowledge, experience and willingness to act in his best interests. Assuming the physician is eager to attract the patient to his practice, he may be tempted to exaggerate his qualities, or at least not to understate them.

As mentioned in subsection 4.2.2, the medical profession has limited information on the workings and the exact effects and side-effects of many medical technologies. The

individual physician will almost certainly lack this information too and, consequently, may feel uncertain about the right actions. This lack of knowledge may result in a wide variety of practice styles – defined as the physician's set of beliefs about the efficacy of particular forms of care before the care is provided (Folland et al. 1997, p. 216). As long as the care is considered appropriate, a different practice style does not mean the physician is an imperfect agent. However, if certain information is available but the physician is negligent in keeping up to date about his field, the agency relationship is not perfect.

A frequently described problem in the patient-physician relationship is supplier-induced demand. Because he is ill informed, the patient is not capable of judging the physician's actions. As Evans (1984, p. 83) noted, inducing demand is exactly what the physician-as-agent should do as otherwise the ill-informed patient may demand too much or too little goods or services. Hence, the physician should induce demand in a way that serves the patient's interests best. However, according to the thesis of supplier-induced demand, under certain circumstances the physician may be prepared to induce the patient to demand more goods or services than he probably would have demanded if he were as well informed as the physician. Although the GP differs from a secondary care physician in that he specifically has a referral function, he also treats patients by himself. This results in a double role: the physician provides the care after he had decided, or advised the patient about the indicated treatment (Evans 1974). Theoretically, the physician is thus in a position to increase the demand for his own goods or services.

Several approaches have been used and several models have been put forward to describe and analyse the problem of supplier-induced demand (see Labelle et al. [1994] and Folland et al. [1997] for reviews). Studied are, for instance, the effect of physician supply on the volume of medical care utilisation, on physician incomes and on fee levels, the effect of fee levels and changes therein on utilisation, and the effect of reimbursement methods on utilisation (Labelle et al. 1994, p. 350). Folland et al. (1997, pp. 170-178) gave an overview of the models developed in the course of time. In the original model it is supposed that a physician is able to use his discretion to increase the quantity of care demanded. A decreasing number of patients per physician, due to an increasing number of physicians, may drive a physician to induce demand. In case the increase in demand is large enough, a new equilibrium may be found in which the price is higher than the initial price. Examples of other models are the target-income model, the discretion model and the profit-maximising model.

Despite the fact that several models are developed, the problem has not been settled yet. Labelle et al. (1994, p. 351) grouped the criticisms into three categories:

1. A lack of a rigorous theoretic model: models are incomplete or findings are consistent with the models as well as with neo-classical theory.
2. Specification errors in econometric models: important variables are omitted, endogeneity of independent variables is not recognised, or demand is underidentified.
3. Measurement errors: the sample is not representative, or a bias occurred due to the use of aggregate data.

The extent to which physicians induce demand seems to be limited. It is estimated that the elasticity of inducement is 0.1. Doubling the number of physicians results in an inducement effect of ten percent then (see, for instance, Rossiter and Wilensky 1983 and

Cromwell and Mitchell 1986). However, the presence of supplier-induced demand is analysed by measuring the increase in demand due to induction by physicians. This induction is supposed to occur as a reaction to lowered fees or a diminished number of patients per physician. The effectiveness of the agency relationship absent a change in fees or physician stock is not analysed. Hence even without this change factual demand may be higher than the patient would have demanded had he had the same information. Whether this factual demand indeed is higher is difficult to assess.

Grytten and Sørensen (2001) examined whether supplier-induced demand existed for primary care services in Norway. They analysed data for two groups of primary care physicians: contract physicians (receiving fixed grants in addition to patient fees per visit and fees for the provision of specific items of medical treatment) and salaried physicians. They found no evidence for supplier-induced demand, which according to the authors could be explained in two ways. Firstly, contract physicians are motivated by other factors, like professional norms and caring concerns, than pure economic factors. The second explanation is that the regular controls by the Norwegian National Insurance Administration are effective in restraining supplier-induced demand.

Supplier-induced demand usually refers to the problem that the physician induces the patient to demand *more* than he would have if he was as well-informed as the physician. The opposite is also possible though. Then, the physician influences the patient's demand in such a way that *less* goods or services are demanded than probably would have been if the patient were as well-informed as his physician. Here the same measurement problem arises. It is difficult to test whether the demand would indeed be higher if the patient was better informed.

The relationship between patient and general practitioner

Principal:	Patient
Agent:	General practitioner
Good/service:	Diagnostic, therapeutic and prognostic information, prescriptions, referrals and (co-ordination of) medical care
Desired actions:	Proper provision of information and (co-ordination of) medical care, that is, according to the patient's preferences in order to obtain a desired outcome (i.e. a certain health status or state of well-being)
Information asymmetry:	The physician is supposed to have superior information about clinical pictures, possible treatments, effects or side-effects of treatments, et cetera
Conflict of interests:	The nature, quantity and quality of the care the patient is prepared to pay for will depend on the price-quality trade-off he makes. However, for reasons of own interest the physician may want to provide a different type or volume of care. A fee-for-service system, for instance, may drive the physician to provide more care than the patient would have been prepared to pay for had he possessed the same information. On the other hand, arguments like leisure or workload may result in a reduction of the amount of care the physician is willing to provide.

4.3 Patient and third party

4.3.1 Introduction

By ignoring the presence of a third party, one passes over the influence this presence has on the structure and the functioning of the health-care sector. A consequence of third-party financing is that the patient is no longer a solo principal operating more or less independent of others, but that he is now member of a group. The risk-spreading function of the third party makes that the patient, as an insured, is confronted with the interests of other insured. The patient's behaviour within his relationship with the physician has external effects now.

4.3.2 Applicant or insured and third party

Agency theory has also been applied to the relationship between insured and third party. Usually, the third party is viewed as principal then (see, for instance, Arrow 1986). As the agent, the applicant has crucial information about his medical history, his current health status, his life style et cetera – the 'quality' of the applicant – and, therefore, may have a better idea of his future health care costs than the third party. The latter is supposed to have an interest in attracting an applicant with expected claim costs lower than the premium charged for a particular policy, or in adjusting the premium or the policy to the applicant's risk. In each option the third party would like to obtain the information the applicant possesses.

Harris and Raviv (1978) viewed the insurer as the principal confronted with a moral-hazard problem. In case the insured pays a fixed premium – independent of the illness and independent of the costs of care – he faces fewer incentives to prevent the disease from occurring (ex ante moral hazard), but he does face an incentive to increase his spending on care (ex post moral hazard). As the insurer can not observe the insured's actions, he has to draw up a contract that stimulates the insured to behave more efficiently. An obvious way to reduce moral hazard is to make the benefits contingent on the nature and the severity of the illness. Once the illness is diagnosed and a payment is made, the insured bears the risk of overspending. However, it is virtually impossible to appraise the incurred loss due to an illness.

A more common way to reduce moral hazard is cost sharing. Although this will depend on the form of cost sharing – coinsurance, copayment, deductible, no-claim – it may reduce moral hazard as it raises the costs of consuming health care to the insured. In the RAND Health Insurance Experiment it has been shown that a reduction of the extent of the coverage leads to a reduction in the amount of care demanded (Newhouse et al. 1996).[9]

[9] See Newhouse et al. (1996) and Bakker (1997) for further details of cost sharing and the RAND experiment.

Another way to reduce moral hazard is to restrict the access to certain providers or the entitlement to relatively expensive care, which the insured may want to substitute for comparable but cheaper care.

The roles of principal and agent may also be inverted so that the applicant becomes principal and the third party becomes agent. In a health care system with a choice of third parties, the applicant faces an adverse-selection problem. The applicant would like a third party that will look after his interests, but it is the third party that possesses information about his performance of the insurance function, the agency function, or the access function – the 'quality' of the third party. To facilitate the selection, the applicant may use information acquired from relatives or acquaintances, consumers' associations, annual reports et cetera.

Once the selection of a third party has taken place, the former applicant may face a new problem, namely the problem of moral hazard. As the insured cannot (perfectly) monitor the insurer's actions, he does not know whether the outcome is the best possible outcome and how far it is the insurer who is responsible for it. A distinction can be made here between two of the different functions a third party may perform: the insurance function and the agency function. In performing these functions, the third party is supposed to act on the insured's behalf. Firstly, he should act as a prudent insurer. Secondly, in performing the agency function, he is supposed to take measures to reduce the problems of adverse selection and moral hazard the insured-as-patient may have within the relationship with his GP.

In case the same party performs both functions, conflict of interests may occur. The outcome that should be attained by performance of the insurance function is risk bearing and payment of claims against the lowest possible premium, given certain coverage. The expenditures should be restricted then. In case of the agency function, however, the third party has to arrange high-quality care. Although high-quality care may result in lower costs, such care may be cost enhancing as well.

The use of controlling and monitoring methods may stimulate the third party to act in the insured's interests. Further, in a competitive health insurance market and under the assumption that the insured is not a 'bad risk' (expected costs higher than the contribution the insurer receives for this particular insured), the insured may use the so-called threat of exit. If the insured is indeed a 'bad risk', then the potential exit of this particular insured will probably not be perceived as a threat to the insurer.

The relationship between applicant or insured and third party

Principal:	Applicant/insured
Agent:	Third party
Good/service:	Reduction of uncertainty, information about providers of care, arrangement of and access to cost-effective high-quality care
Desired actions:	Gaining of information about the insured's preferences, reliable insurance products, proper provision of information about the quality of health-care providers, quick payouts, product development, arrangements with intermediaries and health care providers, reduction of moral hazard, et cetera
Information asymmetry:	The third-party agent has private information about its own behaviour and efforts, about the providers of care, about the behaviour of other insured, et cetera
Conflict of interests:	The third party may have no direct incentive to reveal the above-mentioned information to the insured and may, for instance, have other interests regarding the arrangements with providers of care, the quality of care provided to the insured, the trade-off between quick payouts and receiving interest

4.4 Third party and general practitioner

4.4.1 Introduction

What differentiates the relationship between a third-party agent and a GP from a standard agency relationship is that it is embedded in the triangular relationship between patient, physician and third party, and that it is difficult to view it apart from the patient-physician relationship. Contrary to the latter relationship, which can be viewed as a separate entity, the relationship between third party and physician only exists because of the patient-physician relationship. Hence, the findings of section 4.2 are of use here.

Several authors applied agency theory, more or less extensively, to the relationship between a third party and a physician (see, for instance, Dranove and White 1987, Blomqvist 1991, Robinson 1993, Pontes 1995, Propper 1995b, Smith et al. 1997, Jelovac 2001). Scott (1996) also used agency theory to describe various relationships in health care. He modelled health care as a four-party system in which the government is the third-party payer and the medical profession as a group – a professional body or trade union – is positioned between the third party and the individual physician. In such a model it is recognised that negotiations about contracts may be conducted by, for instance, a professional body. However, the possibility of applying selection, controlling, or monitoring techniques within the relationship between a third party and an individual physician is omitted in the model of Scott. In this thesis, not the negotiations as such are in question, but the resulting arrangements between third party and GP that may influence the behaviour of the physician.

4.4.2 Application of agency theory

The third party is depicted here as the ill-informed principal and the physician as the well-informed agent. The physician has information about his knowledge and experience, his intentions, his own actions et cetera. The third party is supposed to know what should be accomplished but may have difficulties checking whether the physician concerned is the most capable physician, whether he performs the right actions and whether he makes an effort.

Naturally, the exact objectives of both parties are unknown and can only be inferred. The physician will at least take the preferences of the patient into account while making health care decisions. As perfect agent he should even fully act in the patient's best interests. As to the third party, it has been noted that one should identify the nature of the third party. The third-party payer will have different objectives than the third-party agent. Further, the control over a third party may be held by different interest groups. Who controls the third party will determine its objectives (Pauly 1988b). But irrespective of who controls the third party, it is plausible that the objectives of third parties, insured and physicians will diverge.

The interests of a third party only performing the insurance function will be different from those of a third party (also) performing the agency function. While performing the *insurance function*, the third party will probably aim at reducing the risk that its expenditures exceed its revenues. The discipline of the market may force third-party payers to pursue their primary objective, i.e. cost reduction or profit maximisation. This may, for instance, be accomplished by curtailing spending on sales, premium collection, verification and payment of claims, as well as by pooling risks and diversifying (Pauly 1988b) or by charging higher premiums. In performing this function, the third party doesn't necessarily have to be involved in a relationship with the physician. Conflict of interests may even be entirely absent, although the third party may have an interest in a physician economising on goods and services as this may reduce the number of claims.

Performance of the *agency function*, however, entails different interests. This function carries the representation of the insured's preferences. It is true, performance of the insurance function is also in the insured's interests because of a welfare gain due to risk reduction. Moreover, proper performance of this function may result in relatively low costs, which may be translated into a relatively low premium. Looking after the insured's interests by performing the agency function, however, requires the minimisation of agency problems within the patient-provider relationship. Whether performance of this function will result in a conflict of interests between third party and physician, depends on whether both act as perfect agents on behalf of the insured or patient respectively. The size and the nature of a possible conflict will depend on the private preferences of the parties involved.

If the third-party agent performs its function properly, the physician will be stimulated to act as an agent for the patient. Interests may conflict here as the physician may have his own objectives, which may differ from the third party's. Blomqvist (1991, p. 412) pointed at the physician's role as 'double agent' – the patient's as well as the insurer's agent – '(...) which is due to the combination of *both* information asymmetry *and* third-

party financing (...).' Clearly, both roles will conflict once the physician and the third party pursue their own interests instead of the patient's. If both the physician and the third party would act as the patient's perfect agent, then the conflict would be absent for the interests of both parties would reflect the interests of the patient. It should be noticed that the better the physician acts as the patient's agent, the less necessary it is that the third party performs the agency function on behalf of that patient, that is, as far as it concerns the reduction of moral-hazard problems. The function may be beneficial to the patient though, as in the course of time the third party may have acquired information about the quality of several physicians. This information may help the patient reducing the costs of searching the most eligible physician.

The kind of outcome a third party aims at will be related to the performed functions. Is it the insurance function, then it may be expected that within the insurer-physician relationship economising on health care expenditures is what interests the principal most. The outcome is financial by nature then. The insurer-as-agent should aim at an outcome congruent with the outcome the insured-as-patients aim at. Here the outcome is financial as well as non-financial. It is non-financial as it concerns the quality of care and will probably be stated in terms of a certain health status or some state of well-being. As it will be a trade-off between quality and price, the outcome will also have a financial component.

Contrary to what was assumed in section 4.2, it is not the patient but the third party that is supposed to draw up the contractual arrangements to control the agent and to negotiate a compensation scheme.[10] Hence, it is the third party that may face difficulties interpreting the outcome and assessing whether the physician performed the right actions and whether he performed them right, and thus deciding how the physician should be compensated – or even whether he should be compensated at all. Whatever the outcome should be, the third party will be confronted with outcome uncertainty. Again it is uncertain whether the outcome is the result of the physician's actions or external factors, like the natural course of the patient's disease and other factors that influence the patient's health but over which the physician has no control.

4.4.3 Modification of standard assumptions

Some differences between the present agency relationship and the standard agency relationship have already been noticed and relate to the triangularly relationship between patient, GP and insurer. Another difference between the present and the standard relationship is the *conflict of interests*. Again it is the professional relationship between the patient and the physician that causes this difference. In the standard theory it is assumed that the agent behaves opportunistically. However, opportunistic behaviour of the physician within his relationship with the third-party agent may result in opportunistic behaviour within the relationship with his patient. As it may harm the patient or his own reputa-

[10] Often, the negotiations will be conducted by national associations of insurers and physicians. Here it is assumed that the individual parties have the freedom to negotiate at least part of the arrangements by themselves.

tion, the physician may be reluctant to behave in this way. For instance, due to the presence of insurance, the physician may be tempted to cheat and prescribe an unnecessary treatment in order to extract more money from the insurer. The insurer will sustain a financial loss then, but the patient may bear costs too. The short-term effect for the patient may be a financial loss due to cost sharing as well as a physical injury. A minor long-term effect may be the upward pressure on the insurance premiums. Hence professional ethics may refrain the physician from cheating within his relationship with the third party. This does not alter the fact that there remains room for self-interest seeking behaviour though. Due to the 'zone of uncertainty' (see subsection 4.2.2) a physician has room for discretionary behaviour and thus to pursue his own interests. Extra diagnostics or treatments can be prescribed without facing ethical problems. Some margin for opportunistic behaviour remains too. The physician may send in claims for treatments never provided, send in claims twice et cetera.

On the other hand, the physician's opportunism may be beneficial to the patient. Prescribing a fully covered expensive treatment instead of a cheaper but partly or not covered alternative may indeed benefit the patient. The third party, however, will be better off if the physician chooses the cheaper one (although depending on the marginal effect of the expensive treatment, in the long run the patient may be better off too because of the premium effect). Obviously, the third party acts as an imperfect agent here as otherwise it would have preferred the physician's action that benefits the patient most. A major cause of this imperfection is the fact that the insurer is agent on behalf of many principals. The presence of *several principals* forms another difference between the present and the standard agency relationship. Thus far it is suggested here that within his relationship with the physician, the third party will represent the individual insured. However, once the insured consumes care, the third party can not be a perfect agent on an individual insured's behalf. It has to represent a group of insured then and, therefore, has to divide premium revenues among them. This may again result in a conflict of interests. In case of health-care consumption, the individual insured faces low or zero costs against positive benefits. The costs of the consumed care are at the expense of the group of insured although the benefits are less obvious. Therefore, the group will adopt a critical attitude towards the insured's consumption. Some positive benefits may be reaped due to altruistic and egoistic preferences, but the group will be reluctant to pay for excessive use of health care. The third-party agent, therefore, should act according to the preferences of the individual insured under a 'veil of ignorance', that is the insured states his preferences not knowing whether he will become sick or not (ex ante preferences).

Likewise, in performing the agency function, the insurer should take measures to improve the quality of the provided care by reducing the agency problems the insured faces within the patient-physician relationship. The costs of these measures will be borne by the third party, i.e. the group, and will therefore appear low to the individual insured. In performing the agency function, the third party will thus balance the insured's improved benefits against the costs of the measures necessary to gain this improvement. Hence if the physician acts in the interests of the patient, then both may encounter a conflict with the third party as their demand for care may be larger then the amount the group is willing to pay for. The third party may then prompt the physician to consider the interests of

the other insured. The resulting agency problems within the patient-physician relationship are subject of section 4.5.

4.4.4 Agency problems

As in other agency relationships, both the problems of adverse selection and moral hazard may occur within the relationship between third party and physician. The third party may face difficulties assessing the physician's quality, as it is the physician who has the information about his knowledge and experience, his intentions et cetera. The physician has no incentive to reveal this information or may be tempted to exaggerate it. The inclination to exaggerate competence, however, may be limited due to the presence of malpractice legislation. The threat of malpractice suits may keep the physician from entering into a contract that requires him to practice beyond his ability. Moreover, it may be difficult to obtain malpractice insurance for practising care one is not qualified for. To reduce the information-asymmetry, a third party may try to assess the quality of the provider by checking academic qualifications, membership of professional associations, references and state licences (Pontes 1995, p. 59). Further, the third party may induce the physician to reveal his quality by offering him different contracts and letting him choose (self-selection). Jack (2005) argued that different methods of payment can be an efficient way of sorting physicians according to underlying characteristics, like altruism, risk aversion, or ethical standards. Risk averse physicians, for instance, may prefer being paid on a fee-for-service basis.

Adverse selection is mainly a problem in a model with competitive third-party payers offering several contracts from which the physicians can choose. In a system in which just one type of contract is available, the problem of adverse selection will probably be less severe.[11]

A well-described moral-hazard problem is that of supplier-induced demand. As noted in section 4.2, due to his superior knowledge the physician is able to induce the demand for health care goods and services. This may happen in a way that is not in the patient's interests, but merely in the interests of the physician himself. A third-party payer will be affected by the physician's actions, as it has to pay the claims. A third-party agent not only has to pay the claims, but may also face the consequences of the improper level of care provided to the insured on whose behalf it acts. Demand inducement may, for instance, lead to dissatisfaction of the insured about the third-party agent's actions or may have negative effects on the insured's health status.

[11] Even in a National Health Service with one national contract adverse selection may occur. For instance, the contract for GPs may be unacceptable to excellent or even reasonably good physicians (who may succeed in becoming a medical specialist) but may very well be acceptable to inferior physicians.

The relationship between third-party agent and general practitioner

Principal:	Third-party agent
Agent:	General practitioner
Good/service:	Diagnostic, therapeutic and prognostic information, prescriptions, referrals, and (co-ordination of) medical care
Desired actions:	Proper provision of information and (co-ordination of) medical care according to the preferences of the patient involved, but taking into account the interests of the other insured while making health-care decisions
Information asymmetry:	The physician is supposed to have superior information about the clinical picture of the patient involved, the indicated treatment, his own effort et cetera; outcome uncertainty is a complicating factor in observing the physician's actions
Conflict of interests:	The exact conflict depends on the weight put on the different interests: by the third party on the interests of the patient/insured involved, of the other insured, and of its own; by the physician on the interests of his patient, and of his own

4.5 The patient-physician relationship reconsidered

The presence of a third party is of influence on the relationship between a patient and his physician. Firstly, a third party may have increased the *information* both the patient and the physician possess due to the performance of the agency function. The better the patient is informed about medical care and about the supply side, the less problems he may face in selecting a physician and in checking whether the physician performs the right actions and whether he performs them right. The better the physician is informed, the better he is able to serve as the patient's agent.

Secondly, the *interests* of both parties may have been altered, which may result in modified behaviour. It has been argued that the conflict of interests may be less severe than in a standard agency relationship due to, for instance, professional norms and values. On the other hand health insurance provides the physician with the opportunity to pursue his own interests. The physician's altered behaviour due to the presence of insurance may result in an effect often called supplier-induced moral hazard.[12]

The patient's behaviour may be altered too. If the patient is fully insured, then the necessity to make a price-quality trade-off is removed. As long as the benefits of additional care are positive, the patient faces an incentive to demand this care as its costs seem zero at the moment of consumption (abstracted from, for instance, cost sharing and time costs). Actually, the costs are positive and spread over the group of insured.

[12] The term supplier-induced moral hazard is used to denote a specific form of supplier-induced demand: the physician induces demand in the knowledge that the patient's insurance policy covers the resulting costs.

A third remark is about the *contractual arrangements*. In general, it is not the patient who draws up the arrangements that should motivate the physician-as-agent to act in the patient's best interests. At least parts of the arrangements are the result of negotiations between third parties and physicians (or their representatives). Examples are the negotiations for remuneration systems and the level of the payments on which the individual patient mostly has no influence. As a result, the patient is in want of some instruments that might be suited to guide the physician. In the present triangular relationship, the use of such instruments is delegated to the third-party agent. This is different from what is usually supposed in the theory of principal and agent. Furthermore, as the contractual terms are set, it makes less sense for the physician to prefer one contract to another. Although some variation in contracts may remain, like public and private contracts or contracts adjusted for risk, the variation will probably be smaller than if patients were free to devise a contract.

Given that there is not a single best contract, the question arises on whose behalf the physician-as-agent will act. It is unlikely that he will focus on just one patient. Firstly, the contract will not provide him with the financial means to act as an agent on behalf of just one patient. The physician will have to be an agent for other patients too and thus will have to divide his time, energy and means among them. Secondly, as the patient is insured, he faces no costs at the point of service and, therefore, may be tempted to demand unconditional attention, devotion and care. Scarcity of resources may prompt the physician – or he may be encouraged to do so by a third party – to take the interests of other insured, or even of society, into consideration by constraining the demand.

While describing his view on perfect agency, Mooney (1994, p. 94) distinguishes between perfect agency from the viewpoint of an individual patient, from the viewpoint of a physician's patients – that is, the practice population as a whole – and from the viewpoint of society. From the viewpoint of the physician's patients, the perfect agent maximises the group's utility, given constraints like scarce resources. Mooney expects health, information, involvement in the decision-making process, and several aspects of the decision-making and treatment process to be arguments in the patient's utility function. From the viewpoint of society, perfect agency means maximising society's utility, which may consist of efficiency and equity. Mooney excludes the perfect-agent relationship between an individual patient and the physician from further analysis as '(...) the idea of the 'index' patient getting unrestricted and unconstrained treatment based on the ethical principle that the doctor must do the best she can for the patient, irrespective of what else the doctor might be doing, is not an economic issue since it denies the scarcity of resources' (Mooney 1994, p. 94).

Clark and Olsen (1994) consider the physician as an agent facing a budget constraint. Their physician-as-perfect-agent is the decision-maker who has to find a level of equilibrium between maximising the patient's utility (which may be gained from the process of health care as well as the outcome of this process) and the utility of 'society' (a group consisting of patients and potential patients who contribute to the budget).

The relationship between patient and general practitioner revised

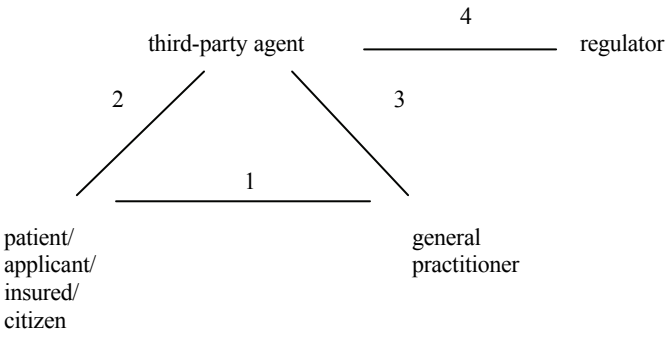
Principal:	Patient
Agent:	General practitioner
Good/service:	Diagnostic, therapeutic and prognostic information, prescriptions, referrals and (co-ordination of) medical care
Desired actions:	Proper provision of information and (co-ordination of) medical care, that is, according to the patient's preferences in order to obtain a desired outcome (i.e. a certain health status or state of well-being)
Information asymmetry:	The physician is supposed to have superior information about clinical pictures, possible treatments, effects or side-effects of treatments, et cetera
Conflict of interests:	Now that the patient is insured, the necessity to make a price-quality trade-off has been removed. As long as the marginal benefits outweigh the residual costs (like time costs) the patient has an incentive to demand for care. The physician may not just direct his actions at one single patient, but may take the interests of his other patients, or even of society into account while deciding upon his actions.

4.6 Theoretical framework of the agency relationships

How can agency theory be used to analyse and compare the different type of (existing) relationships between third-party agents and general practitioners?

The combination of the relationships discussed here results in a theoretical framework that serves as the background to the further analysis of the relationship between third parties and GPs. This framework consists of a set of four agency relationships (see figure 4.1).[13]

Figure 4.1. Four agency relationships



[13] Some relationships are omitted here. For instance, the regulator may engage in direct relationships with GPs (licensure, quality controls et cetera).

The first agency relationship: between patient and general practitioner
The first relationship considered here is the one between a patient (the principal) and a GP (the agent). Due to the asymmetric information and the physician's own interests, the physician may act as the patient's imperfect agent. As a result, the patient may encounter problems of adverse selection and moral hazard.

The second agency relationship: between applicant or insured and third-party agent
For several reasons, the same principal (now in his role as applicant) may want to enter into a relationship with a third party. One reason may be the reduction of his financial risk. Another reason may be the reduction of agency problems the applicant (that is, the patient) may face within his relationship with the physician, in which case entering into a relationship with a third-party agent may help. But then, he may have difficulties selecting the third-party agent who will serve his interests best – this under the assumption that there are several third parties to choose from.

Within the framework employed here it is assumed that both the third-party agent and the GP pursue also their own interests and thus act as imperfect agents for the patients. Measures should be taken then to try to control both parties. Although one could think of other parties fulfilling this role, the third-party agent is viewed here as the principal to whom the design of contracts is delegated and that has to control and monitor the GP. But having selected a third-party agent, the insured faces difficulties checking whether this agent actually performs his agency function properly. The insured may try to control and monitor the third-party agent. Further, in a competitive insurance market the insured is, in principle, in a position to choose another third-party agent and may therefore use the threat of exit as an incentive (i.e. deselect the agent).

The third agency relationship: between third-party agent and general practitioner
In performing the agency function, the third party is supposed to take measures to reduce the problems the patient encounters within his relationship with the GP. Therefore, the third-party agent should, as a principal, enter into an agency relationship with the physician concerned. This is the third agency relationship.

The fourth agency relationship: between regulator and third-party agent
Despite of the presence of techniques like (de)selection, control and monitoring, it is generally recognised that the insured is in a weak position. This explains the presence of a fourth agency relationship: between the third-party agent (the agent) and a second third party, i.e. a regulator (the principal). This regulator, for his part, is supposed to take measures to reduce the problems the insured may have within his relationship with the third-party agent and to stimulate the third-party agent to act in the insured's best interests. Thus, while trying to remove the imperfections in the third-party agent's functioning as agent on behalf of the applicant or insured, the regulator acts here as another agent.[14] It

[14] This regulator may, for instance, be a government. It will be obvious that this creates another set of agency problems with the applicant, insured or patient as principal and a government, for instance chosen by means of voting, as agent.

is assumed here that the regulator actually uses instruments to effectuate that the insurer will indeed act as the insured's agent. The regulator may, for instance, further managed competition between insurers, institute supervision, or provide information.[15]

To be able to fulfil its role as the insured's agent properly, the third-party agent has to gain information about the insured's preferences first. Having acquired this information, it has to take measures to promote a relationship between patient and physician that approaches a perfect agency relationship. The physician is agent on behalf of the rest of his practice population too and, furthermore, will be encouraged by the third-party agent to take notice of the other insured. 'Perfect' implies acting according to the patient's ex ante preferences (under a veil of ignorance). The third-party agent will thus aim his attempts to improve the agency relationship between the physician and the individual patient especially at the physician's own interests by, for instance, increasing the compatibility of incentives.

Starting from standard agency theory, one might expect to find a third-party agent controlling and monitoring a selected group of physicians. However, Arrow (1986, pp. 1193-1194) mentioned some reasons why the arrangements found in practice may vary from the theoretical arrangements. Such reasons are transaction costs or the existence of professional norms (see also section 3.4). Further, in the preceding pages it has been argued that relationships in the health care sector deviate from a standard relationship, which is for an important part the result of overlapping interests. In addition, instead of a third party the regulator may have assumed the agency role, leaving the third party to fulfil only the insurance function. Hence the question is whether, and to what extent, the three main strategies (selecting, controlling and monitoring an agent) are actually found within these relationships. In the following chapter a closer look is taken at the several instruments third-party agents may and actually do use.

4.7 Summary and conclusion

Agency theory is not specifically aimed at relationships in health care. Nevertheless, in view of the characteristics of the relationships analysed in this chapter, its use here seems justified. Regarding the relationship between patient and physician, some authors have even pointed at it as an obvious example of an agency relationship. Although less prominently, also the relationship between a third party and a physician has been described in agency terms.

Health-care relationships are not straightforward agency relationships. In several respects they differ from the standard agency ones. First of all, in relationships in which a GP is involved, the conflict of interests may be relatively small. This is because of professional norms and due to the fact that as a professional the physician will include at least part of the patient's interests in his own objectives.

[15] The questions which measures should be taken to control the third-party agent and which parties are the obvious ones to take these measures are beyond the scope of this research.

Secondly, the relationship between a patient and a GP usually is a lengthy or repeated one, so poor performance may prompt the patient to terminate or not to renew the relationship and may ruin the physician's reputation.

Thirdly, due to his superior knowledge and experience, it will often be the agent (i.e. the physician) instead of the principal (i.e. the patient) who specifies the goals that are ultimately achievable. Moreover, often these goals can only be set after the patient has been diagnosed and thus after the relationship has been established. It should be noticed that the patient as the ill-informed principal visits a professional because of his superior knowledge and experience. Because the patient can not be an expert in all fields himself, he expects that the physician indeed is a professional and thus expects that the physician has superior knowledge and experience. Paradoxically, the patient thus enters an agency relationship expecting and hoping for the physician's information surplus.

Although it is exactly the informational advantage that is an important reason for visiting a professional (i.e. the physician), this advantage gives the physician the opportunity to induce the patient's demand for his goods or services. Although inducing demand is what a physician-as-agent should do, the physician may be able to persuade the patient to demand more goods or services than the patient probably would have demanded if he were as informed as his physician is. If the patient has health insurance, then a special form of supplier-induced demand, namely supplier-induced moral hazard, may occur.

This brings us to the final difference, namely the presence of a third party besides the first (i.e. the patient or the insured) and the second party (i.e. the physician). These parties form a triangular relationship. The presence of a third party may have an influence on the information first party and second party possess as well as on their interests (for instance, because of the altered financial incentives they face). Also, what happens in one relationship within the triangle is probably of influence on the other party. The presence and role of a third party has some additional effects, like the presence of others' (i.e. other insured') interests and thus of external effects. Another effect is that now it will be the third party instead of the patient who will be the main designer of the contractual arrangements with the physician and who will use the several instruments to promote that the physician acts in the patients best interests. This leaves the patient as a principal with only a very limited set of techniques to influence the behaviour of the GP.

In spite of the fact that health-care relationships differ from standard agency relationships, the agency characteristics and the resulting agency problems are obviously present within the relationships described here. Hence the research question of this chapter:

Is agency theory applicable to relationships in the health care sector in general, and to the relationships between third-party agents and general practitioners in particular?

is answered positively.

By means of the agency concepts, we constructed a theoretical framework consisting of a set of agency relationships and providing the rationale for the use by a third party of the several strategies in order to influence the GPs. The set consists of four agency relationships.

The *first* relationship is the one between patient (the principal) and GP (the agent). Because of asymmetric information and diverging interests, the patient may encounter problems of adverse selection and moral hazard. In order to reduce his financial risk associated with health care use and to reduce the agency problems the patient may face within his relationship with the physician, he may want to enter into a relationship with a third-party agent; the *second* agency relationship. But then, he may face difficulties in selecting the third-party agent who will serve his interests best.

Once selected, the third party as an agent for the insured is supposed to take measures to reduce the problems the patient may encounter within his relationship with the GP. To be an effective agent, the third-party agent will have to enter into a relationship with the physician; the *third* agency relationship. Whether the third party is indeed an agent that serves the insured's interests, is also a part of the problem within the *second* agency relationship.

The remaining agency relationship mentioned in the description of the framework is the relationship between the third-party agent and the so-called regulator (the *fourth* agency relationship). In spite of the several techniques available to the insured, like (de)selection, control and monitoring, the insured's weak position creates a need for a so-called regulator. The regulator will have to reduce the problems the insured may have within his relationship with the third-party agent and to stimulate the latter to serve the insured's interests. It may, for instance, further managed competition, institute supervision or provide information.

Within the theoretical framework, it is assumed that the insured as well as the regulator stimulate the third-party agent to gain information about the insured's preferences and to take measures in order to improve the agency relationship between patient and physician. One might expect then to find a third-party agent controlling and monitoring a selected group of physicians. For several reasons though (like the presence of positive transaction costs and professional norms), the factual arrangements will probably differ from the theoretical arrangements. Moreover, it has been argued that because of overlapping interests the relationships in health care may deviate from a standard agency relationship. Hence the question is to what extent the three main strategies (selecting, controlling and monitoring) are actually used by third-party agents within their relationships with GPs. This is subject of the following chapter.

5 AGENCY AND MANAGED CARE

5.1 Introduction

The theoretical framework constructed in the previous chapter is applied here to analyse whether third-party agents actually use the several strategies in order to influence the behaviour of GPs. A literature study is conducted to investigate which strategies third parties (like GP fundholders in the UK, and managed-care organisations in the US) may or do employ and with what effect. The research question central to this chapter is:

Which techniques can and do third-party agents apply within their relationships with general practitioners in order to reduce the agency problems within the patient-physician relationship?

In health care (the use of) such a set of techniques is often designated as 'managed care'. In section 5.2 we will analyse this rather broad but also rather widely used concept in terms of agency theory, identify the several phases of managed care and bring these in connection with each other. Next, we will identify several techniques that are applied in health care, and group them according to the triptych of agency theory (sections 5.3, 5.4 and 5.5). In section 5.6 we will briefly discuss some issues of the use of managed care and the difficulties of drawing conclusions from the managed-care literature.

5.2 Managed care and agency

5.2.1 A definition of managed care

The techniques used by third-party agents in health care are often grouped together under the term 'managed care'. Although the term 'managed care' is widely used, there is no consensus about its precise meaning. As a result, the definitions often differ from each other. Glied reviewed the managed-care literature extensively and noted that there is no single definition of the term managed care that is broadly accepted. Because of the tremendous variation in the nature of the managed-care plans, it is difficult to assess the economics of managed care theoretically as well as empirically. Hence another way to look at managed-care plans is to think of these plans as combinations of various sets of mechanisms. These mechanisms have changed over time though. According to Glied, economic theory and empirical research have not kept pace with the development of managed care and research is needed to identify the characteristics of managed care that generate economically meaningful differences in outcomes (Glied 2000).

When it comes to managed care, authors seem to have different opinions about at least four matters. Firstly, the term is often used to refer to several techniques, but sometimes it

is used to denote the diverse organisational arrangements within which these techniques are employed (Weiner and de Lissovoy 1993).

Secondly, there is disagreement about the several techniques as such. The question is then whether the use of a certain technique is essential for the notion of managed care, and whether can be spoken of managed care if just one technique is applied or only if at least several techniques are used. This may be illustrated by the following definitions. Miller and Luft (1994, p. 1512), for instance, restricted the term by stating that 'The selection of network physicians is the starting point for such management and is the single most important feature that distinguishes a managed care from an indemnity (fee-for-service) plan with utilisation management.' In other words, selection of physicians is an essential part of managed care and differentiates managed care from non-managed care. Just like Miller and Luft, Weiner and de Lissovoy (993, p. 89) used an 'at-a-minimum definition' of managed care. Instead of selection, however, they viewed pre-admission certification (pre-admission review) as '(…) the element that distinguishes (…) managed care plans from non-managed indemnity plans.'

A third group of definitions of managed care focuses on the goals that should be accomplished with managed-care techniques. According to Rodwin (1995, p. 604), 'Managed care changes traditional indemnity insurance and fee-for-service practice by integrating the financing and delivery of medical services, with the aim of *controlling costs and improving quality*' (italics added, A.V.). Iglehart (1994, p. 1167), on the contrary, stated that the term refers to '(...) a variety of methods of financing and organising the delivery of comprehensive health care in which an attempt is made to *control costs* by controlling the provision of services' (italics added, A.V.).The definitions agree with each other on the fact that managed care is about the financing as well as the organisation of health care. A major difference, however, is that according to the second definition the goal of managed care is cost control, whereas in the first equal attention is paid to the quality-improving potency of managed care. This difference reflects the divergent views on managed care; as a means to contain costs, as a means to contain costs while at least maintaining quality, or as a way to improve quality (preferably against relatively low costs). As a result of the emphasis some put on the economics of managed care, the term 'managed care' has a negative connotation. It is often associated with cost containment and external control (Ellis and Burns 1998). Research by the Steering Committee on Future Health Scenarios (STG 1997) showed that, in general, Dutch GPs are opposed to managed care. Partly, this seemed the result of ignorance. Further, they were of opinion that managed care puts an extra burden on them but doesn't lead to significant improvements in the quality of care.[1]

Finally, the question is whether it is or should be a third-party payer who applies the managed-care techniques. The aforementioned definitions seem to suggest that it indeed should be. Both Miller and Luft (1994, p. 1512) and Weiner and de Lissovoy (1993, p. 89), for instance, spoke of managed care plans and indemnity (fee-for-service) plans. Iglehart (1994, p. 1167) and Rodwin (1995, p. 604) considered managed care as an integration of financing and organisation of health care. However, a group practice or an as-

[1] See STG (1997) for opinions of other parties on managed care.

sociation of GPs may as well function as third party that may try to influence an individual physician's decision-making. In that case, managed care can not be considered an infringement on the professional autonomy, i.e. the autonomy of the profession as a whole. At most, an individual physician's autonomy is reduced then. An example of a definition in which an infringement on the autonomy of patient and physician is displayed is the definition by Van der Werf (1997, p. 512) who interpreted managed care as a way of directing care whereby a third party appropriates responsibilities and power within the physician-patient relationship.

Managed care is regarded here as the attempts of a third party, by means of one or more techniques, to influence the decision-making process within the relationship between an individual patient and an individual provider of care, i.e. the GP. These decisions are no longer the exclusive domain of patient and physician. If this decision-making is considered to be affected by patients as well as by physicians, then a third party wanting to manage the care by influencing this decision-making has two options. The first option is the application of the managed-care techniques within the relationships with physicians, as is described in this chapter. A second option would be to try to manage care via the insured by:

- *selecting* potential insured (cream skimming);
- *controlling* them by means of cost sharing (co-payment, deductible, co-insurance, or no-claim discounts), policy conditions, consumer information et cetera;
- and by *monitoring* them, for instance, through analysing claims or checking preventive measures to be taken by the insured.

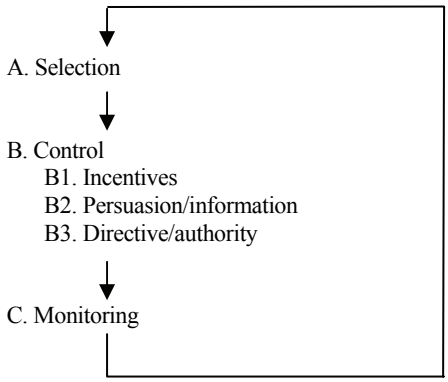
In fact, such strategies to influence the behaviour of the insured may also be considered managed-care techniques. However, while speaking of managed care authors usually refer to techniques applied within the relationships between third party and physicians. This may reflect an economic view of the patient-physician relationship in which the physician is considered the patient's agent having superior knowledge and experience and to whom, therefore, patient care decision-making is delegated. In medical sociology the patient has a more prominent role. Although depending on variables like the nature of the disease or the patient's socio-economic status, also the patient contributes to the decision-making. Decision-making is then considered as a bargaining process. In this thesis managed care is used in a more narrow sense, focusing on the techniques used in the relationships between third-party agents and GPs.

5.2.2 The application of agency theory to managed care

From the perspective of agency theory, managed care can be viewed as a set of techniques that a principal may employ with the purpose of influencing the behaviour of an agent in a way beneficial to the principal. If it is assumed that the interests of the third party (the principal) are also the patient's interests, then the ultimate purpose of managed care is to encourage the GP (the agent) to act in the patient's interests.

The triptych of agency theory can be applied to classify the various managed-care techniques. It consists of three successive phases that form an iterative process. Likewise, managed care can be viewed as an iterative process (see figure 5.1).

Figure 5.1. The managed-care cycle



The first step of the managed-care cycle, *selection*, is made to identify and contract those GPs of whom it may be expected that they will provide high-quality care in a cost-effective way.

Once physicians are selected, a third-party agent may strive to make contractual arrangements about the provision of care; the second step in the managed-care cycle. The third party has several *control* techniques at its disposal then, like financial incentives, practice guidelines, physician profiling, and a group of techniques often labelled as utilisation management.

The third step of the managed-care cycle is *monitoring* to check whether, as far as observable, the outcome is satisfactory. The problem of outcome measurability may be evaded by trying to observe the physician's behaviour. As the outcome will only partially result from the physician's actions, an alternative may be to monitor the other outcome-influencing factors and, by so doing, determine the physician's contribution. The third-party agent can use the assembled information in deciding upon 'reselecting' or 'deselecting' the GP (i.e. to extend or to renew his contract, or to terminate his contract respectively).

After monitoring the GP, the managed-care cycle is completed and may be run through again. The next step in the cycle, selection, will often be taken implicitly. Monitoring, for instance, may reveal information on the outcome moving the third party to take additional measures. Clinical rules may be modified, or the information may be used to provide the physician with feedback.

The use of managed care gains in effectiveness if the three successive phases of the managed-care cycle, i.e. selection, control and monitoring, are designed and used coherently.

5.3 Selecting general practitioners

5.3.1 Introduction

A first step third-party agents may take in an attempt to manage the health care on behalf of their members is the selection of GPs. By physician selection is meant here that a third party selects and contracts a part of the total physician stock. The selected group may range from only one physician to a hundred percent of the total number of eligible physicians.

Selection is considered to be an important aspect of managed care (Langwell 1990). Martin et al. (1985, p. 74) suggested that selection of providers may even be the most important feature in a cost containment program. Miller and Luft (1994, p. 1512) also recognised the importance of physician selection and viewed it as the starting point for and the most crucial feature of managed care. There is a large difference in practice behaviour among physicians in different health care settings, but even within a certain setting there is often a wide variation. Luft (1981, p. 138) suggested that this variation indicates that there is no 'standard practice', and 'that the selection of particular practitioners to join a group may have a substantial influence on the average observed behavior within a setting.'

Physician selection is a common feature of managed-care organisations (MCOs) like Health Maintenance Organisations (HMOs) or Preferred Provider Organisations (PPOs). Although, again according to Luft (1981, p. 138), the variability in practice behaviour does not seem to be less in an HMO than it is in non-managed practices, it may well be the case that such organisations select and attract physicians who have on average a more conservative practice style. A 'conservative practice style' suggests that the primary goal of physician selection is to select and to contract those physicians who act in a cost-effective way, and preferably order less diagnostic tests, refer less patients to specialists, and prescribe less drugs. The goal thus is containment of health care costs. Indeed, the strategy is often viewed as a major cost-containment strategy. Johns (1985), for instance, argued that at least in California the strategy was introduced as such. In 1982, selective contracting was made the preferred policy approach on state level by means of legislation.

Further, in their report of the case study of United Healthcare, Martin et al. mentioned the same argument. United Healthcare was an Independent Practice Association organised by the SAFECO Insurance Company (see also subsection 7.4.2). It started in 1974, became the largest of its kind in the United States and was terminated in 1982 because of major losses. At the outset, United Healthcare strove to attract as many physicians as possible, and no attempt was made to select cost-conscious physicians.[2] Consequently, it also attracted physicians who provided cost-ineffective care. Physician selection was applied only at the end of its history. The lack of the selection technique is viewed as one of

[2] An important reason for contracting as many physicians as possible was that it was considered attractive for potential members if they would not have to change physicians once they would choose for United Healthcare (Martin et al. 1985, p. 56).

the reasons for the failure of the plan. In retrospect, United Healthcare managers viewed initial selection of physicians as the most crucial cost-containment feature for a health plan (Martin et al. 1985, p. 74).

Besides cost containment, the selection of high-quality providers may be another important goal. Starting from the idea of the third-party agent arranging care on behalf of its clients, the selection of providers who are able to meet certain quality standards may be of equal importance. Proponents of selective contracting defend it as a strategy to lower health care costs as well as a means to improve patient care. Moreover, cost containment and high quality may be two sides of the same coin, as in the long run a wrong diagnosis and treatment will be more expensive. It is being asserted that the quality of care as well as the quality of the ones providing such care is very important in contracting decisions, and that this appears from the fact that some managed-care organisations require certain qualifications, like board certification (see Bindman et al. 1998, p. 675).[3] Moreover, the goal of selective contracting seems to have changed in the coarse of time. Was it originally a way to contract physicians (regardless of their way of practising) in order to obtain discounts (in return for patients), later it became a strategy to contract cost-effective and high-quality providers (Robinson 1993).

In some respects, selection of physicians by a third party resembles preferred-risk selection (or cream skimming) by a third party on the health-insurance market. In the latter case, the third party tries to select members for whom the expected health care costs are lower than the premium these members have to pay. Preferred-risk selection may occur before or after an insurance contract is concluded (see, for instance, Van de Ven and Van Vliet 1992). As argued in the foregoing, in case of physician selection the third party may try to select physicians who practice in a cost-conscious way. It may, for instance, select those physicians for whom the (expected) costs are lower than the (expected) revenues.[4] Equally, physician selection may occur before or after a contract between both parties is concluded. It will be easier, though, to select preferred physicians before contracting, than to try to get rid of them or to change their values and behaviour afterwards (Martin et al. 1985). Gold et al. (1995a, p. 1679) conducted a survey to identify the arrangements made between managed-care organisations and physicians. Their research indicated that managed-care organisations indeed prefer to select physicians before a contract is concluded. The majority of the plans (71 percent) chose 'careful selection' whereas only 18 percent chose to 'prune later'.[5]

[3] Having a 'board certification' means that the physician is recognised by a medical board to be qualified to work in his field.

[4] It was also noted, however, that a certain physician may have been selected for quality reasons. Than, the similarity between preferred-risk selection on the health-insurance market and physician selection does not hold. Still, high-quality physicians may be considered good risks as they may attract new insured due to their good reputation. On the other hand, high-quality providers may attract expensive patients, i.e. potentially bad risks (expected health care costs are higher than the premium they have to pay). This will especially hold for some specialised physicians attracting chronically-ill patients.

[5] Note that by 'careful selection' Gold et al. (1995a) meant selection before a contract is concluded. In this thesis, the term 'selection' is being used in two ways: to denote selection of physicians and the possible denial of a contract as well as selection and the possible termination of a contract.

What is labelled physician selection here in fact consists of two phases: the selection phase, and the contracting phase. In the first phase the third party has to select (i.e. to identify) the physicians who meet its requirements. In the second phase the third party actually has to contract the selected physicians.

5.3.2 The selection phase

Third parties may try to select preferred physicians in several ways. A first way is to use claim data, which they may have acquired during a previous contract. Analysing such data may reveal information, for instance about physicians' billing practices and the cost-effectiveness of their behaviour. Problematic is that a large amount of data is required and that this data should be reliable and controlled for case mix (Gold et al. 1995b). A physician who prescribes in a certain time period, for instance, 1.3 times the amount of drugs a colleague prescribes, is not necessarily less efficient than his colleague. Maybe the drugs are a (cheaper) substitute for hospital care. Another reason may be that the practice populations of both physicians differ. Of course, practice size is a crucial determinant here, but also is the case mix of the practice population. Further, the physician may provide better care than his colleague may. Gold et al. (1995a, p. 1680) found that only 37 percent of the plans used quantitative data about physicians' performance and practice style for the selection of new physicians. Sixty-three percent, however, used qualitative data, like professional reputation and patterns of care.

Another way to select physicians is credentialing (Gold et al. 1995b). Credentialing of physicians includes verifying licenses, malpractice histories, hospital privileges et cetera. Gold et al. found that all plans credentialed physicians before concluding a contract. Further, managed-care organisations often visit physicians' offices and check whether the facilities meet certain standards (Gold et al. 1995a, p. 1680). Medical as well as administrative facilities may be reviewed then.

Bindman et al. (1998) investigated which characteristics of primary care physicians and their practices were associated with denials and terminations of managed-care contracts between these physicians and managed-care organisations. They found that contracting by managed-care organisations is indeed done selectively. Eighty-seven percent of the 947 respondents – these were office-based primary care physicians in California – had at least one contract with an individual practice association model HMO or with a direct contract HMO. Twenty-two percent of the physicians had experienced a denial or a termination of a contract with one of these types of HMOs. Bindman et al. argued that this might be a larger proportion than expected given the demand for primary care physicians in managed-care organisations. Of the characteristics, practice size was the strongest predictor of denials and terminations. The smaller the physician's practice, the more likely it was that he was denied or terminated from a contract once. Further, primary care physicians with a large proportion of uninsured and non-white patients were significantly less likely to have managed-care patients within their practice. No association was found between age, sex or race of the physician and denial or termination. Board certification was also not associated with being denied or terminated from a managed-care contract,

but it was twice as likely that certified physicians had at least one contract with a managed-care organisation.[6]

In the United Kingdom, general practices were eligible for fundholding status in case they met a certain list-size as well as administration and automation standards. Further, all partners of the practice had to agree with participation in the fundholding scheme (Audit Commission 1995). Formulating such terms was a way to select those GPs of which it may be expected that they were capable of being a fundholding agent for their patients.

On the one hand, it may be hard for a third party to select physicians who are cost conscious and provide high-quality care, as the physicians' characteristics are difficult or costly to observe. On the other hand, self-selection may make the selection process somewhat less complicated (Martin et al. 1985, Arrow 1986). Careful marketing and stating explicitly in advance that cost-effective and high-quality care is stimulated by the use of managed-care techniques may enhance self-selection.[7] By so doing, it may be expected that the third party will predominantly attract physicians whose ideas about practising health care harmonise with those of the third party. According to Sørensen and Grytten (2003), for instance, more talented and motivated people will try to find and accept positions with performance compensation, while others will seek traditional wage contracts. They tested whether experienced and productive physicians and physicians who prioritize professional work rather than leisure-time and family life tend to prefer fee-for-service contracts, and whether less experienced and productive physicians and those who prioritize leisure-time and family life tend to prefer employment positions with a fixed salary. Indeed their results indicate that the probability of having a fee-for-service contract increases with age and seniority, and that priority given to leisure-time and family life reduces the probability of a fee-for-service contract. Further, Jack (2005) argued that offering the physician different payment methods can be an efficient way of sorting physicians according to underlying characteristics, like altruism, risk aversion, or ethical standards. Risk averse physicians, for instance, may prefer a fee-for-service system, whereas less risk averse physicians may opt for a capitation payment.

Nevertheless, in several cases the assumption of self-selecting physicians may not hold. One reason is that physicians may be forced by local circumstances to apply for a contract with a third party. Firstly, they may apply for such a contract not because they agree with the terms of it, but because of competition for patients. For example, if the majority of patients favour a fundholding GP because they provide or arrange better care, then non-fundholders may be forced to apply for fundholding status too. Secondly, physicians may be forced to apply for a contract because otherwise they would not obtain enough patients. This will especially be the case if a third party has a large market share within the physician's area and if it stimulates its members to use the selected physicians.

[6] Bindman et al. (1998, p. 679) noted that because it is widely recognised that IPAs and direct contract HMOs require board-certified physicians, physicians without certification may not even apply for a contract. As a result, they do not encounter a denial or a termination of a contract.

[7] See also subsection 3.3.2.

Not having a contract with this third party involves missing a substantial group of patients. Another reason why self-selection will not automatically provide the third party with the most suitable candidates is that physicians may lack information about their own behaviour relative to the behaviour of other practitioners. As a result, they may falsely view themselves as cost-effective and high-quality providers.

5.3.3 The contract phase

After the third party has overcome the problems of the selection phase, it actually has to conclude contracts with the physicians. Here is touched on the problems a third party may face then.

Firstly, the ratio of the number of physicians the third party wants to contract to the total supply of physicians should be smaller than one. Does it equal one or is it even larger than one, then the third party is forced to contract all the physicians available, otherwise it cannot guarantee its members access to care. Constraints on the supply of physicians, therefore, limit the usefulness of the selection instrument by third parties.

Secondly, even if the above mentioned ratio is smaller than one, a third party may be forced to contract all GPs who want to do so. By 1994, 21 states in the US had enacted 'any willing provider' laws. Due to these laws, third parties are obliged to contract physicians who are willing to accept its contract terms (Iglehart 1994). Although selection of a certain type of physician is still possible by means of contract design, it is within those states not longer possible on an individual basis.

Thirdly, physicians may object to selective contracting for reasons of solidarity or because it may disturb the network in which they co-operate. Also an association of GPs as a whole may boycott selective contracting and forbid its members to take up an offer individually. In some countries or US states, however, antitrust laws prohibit such cartels.

Finally, the third party as agent has to take account of the members' wishes. Selective contracting only makes sense if patients visit the selected physicians. To encourage the use of these physicians, the third party may stimulate their members financially to visit them in case health care is needed. Probably, the care provided is paid for in full if it is provided by a selected physician. If not, the patient will have to pay a part of the bill (as, for instance, in a preferred provider organisation) or even has to pay the whole bill himself (as, for instance, in an exclusive provider organisation). The consequence of selective contracting a subset of the total supply of physicians will thus be that patients' choice is restricted.

5.4 Controlling general practitioners

5.4.1 Control by incentives

Usually, a GP has to choose an alternative from a set of possible actions. For instance, he has to decide whether or not to treat or whether or not to refer a patient, which diagnostic test he should order, whether he should prescribe generic drugs or brand names, et cetera.

By means of incentives a third party may attempt to influence these choices. The physician is still left a choice and, ideally, chooses the action that is most beneficial to the third-party agent and thus – if the third party performs his agency function well – to the patient.

5.4.1.1 Financial incentives[8]

The use of financial incentives is based on the idea that such incentives influence the behaviour of (primary care) physicians and that this behaviour has an effect of the costs and quality of care. Two basic economic theories are important here (Hsiao 1992). The first one is the demand-side theory. Followers of this theory consider health care to be equal to other goods or services on the market. The demand for health-care goods and services results from decisions by informed and independent buyers of care and depends on their 'willingness to pay'. Suppliers of health care have no or only little possibility to influence this demand or to set their prices independent of it.

A second important theory is the supply-side theory. Followers argue that health care differs from other goods or services. This mainly results from market failures, like the presence of uncertainty (with health insurance and hence reduced price sensitivity as a result) and an asymmetry of information (with the physician as the dominant decision-maker; see also chapter 2).

Within the framework of the supply-side theory fits the assumption that physicians are able to influence the demand for care. Under certain circumstances, like a decreasing number of patients per physician, physicians may exploit this opportunity. Especially if the physician provides the diagnosis as well as the care, then there is room for demand inducement (Van Doorslaer 1988).[9] This is due to the physician's double role as agent for the patient and as provider of care. Hence physicians are able to shift the demand curve, so it cannot considered to be autonomous any longer (Schut 1988). The result is a positive correlation between supply, price and use of care. (See also subsection 4.2.5.)

Obviously, physicians do not always act in their patients' best interests. According to Flierman (1991) there are two explanations for this. Firstly, the physician's utility function contains an income and a leisure variable, besides a medical-ethics variable. The physician may get professional satisfaction from acting in the patients' interests, but at the same time this acting may influence his income and the amount of his leisure time. Secondly, it is not always clear what is in the patients' best interests. This professional uncertainty may manifest itself in uncertainty about the physician's competency or in uncertainty about the moment and the way care should be provided (see also subsection 4.2.2 for the 'zone of uncertainty'). If the optimal treatment is not evident, then the role of income and leisure may increase. In what way the physician's choice of an action out of a set of possible ones influences his income depends on the payment system. The main payment systems are discussed hereafter.

[8] The use of financial incentives in order to influence the behaviour of GPs is also subject of the next chapter. Therefore, this subsection is an introduction to the subject.

[9] Besides the care prescribed and provided by the same physician, one can discern the care demanded by the patient and the care provided by another physician than the physician whom prescribed it.

Two main categories of financial incentives are distinguished here: basic payments and ancillary payments.

Basic payments

The first category consists of basic methods to pay GPs for the services they provide, i.e. for primary care. The predominant methods are fee for service, capitation and salary; case-payment systems are still uncommon. Each method provides a physician with certain incentives and has certain potential effects. Whether certain effects are considered an advantage or a disadvantage depends on one's viewpoint and one's objectives. If, for instance, a third party wants to reduce the number of primary-care services, then the incentives originating from a fee-for-service system may be considered undesirable. However, an attempt to substitute primary care for secondary care may be supported by the introduction of such a system. The fact that in table 5.1 the incentive of a fee-for-service system to provide extra services is considered to be a disadvantage can therefore be disputed.

In case of *fee for service* the physician's income is the product of the number of services and the fees paid for these services. A fee-for-service system stimulates physicians to provide additional services as long as the marginal benefits outweigh the marginal costs of providing them.

In a *capitation* system, a physician is paid a fixed payment per patient for a defined period of time (usually a month or a year) and for a defined package of health care services. Several incentives evolve from capitation. The physician is stimulated to promote prevention and to provide care efficiently, but also to reduce the quality of services (quality skimping), to refer patients (cost shifting), and to select patients on the basis of their expected costs (cream skimming).

In a *salary* system a physician has no incentives to increase the number of services or to select patients. There is no direct incentive to increase productivity or efficiency, or to increase the quality of care. As to the latter aspect on the other hand, the absence of an incentive to increase the volume of care reduces the danger of over-provision of care. Hence a salary system may have positive effects on the quality of care as well.

Jegers et al. (2002) distinguished between fixed versus variable payment systems and between retrospective versus prospective payment systems. In case of a fixed payment system, there is no link between payments and the physician's activities, whereas such a link does exist in a variable payment system. In case of retrospective payment system, there is a link between payments and the physician's costs, whereas this link is absent in prospective payment systems.

Table 5.1 Advantages and disadvantages of basic methods of payment

BASIC PAYMENT	**ADVANTAGES**	**DISADVANTAGES**
Fee for service	- Automatic adjustment for case complexity - Close relationship between remuneration and services - Provides insight into practice profiles - May be applied to stimulate the provision of care - Much used, often preferred by physicians	- Incentive to provide extra services - In case the ratio of fees to effort is not constant, a shift to the more profitable services may occur - Administratively complicated - Difficult to set a budget in advance
Capitation	- Administratively simple - Provides physician with incentive to minimise the costs per case treated - Budgeting in advance is possible - Incentive to provide preventive care	- Risk of patient selection by physician - Risk of lower quality (quality skimping) - Provides physician with incentive to refer (especially expensive) patients - Unfair distribution of means in case the average morbidity in practice populations differs largely - Provides little insight into practice profiles
Salary	- Administratively simple - No influence of profitability of services - Allows for co-operation between physicians - Budgeting in advance is possible - No incentive for cream skimming	- Risk of decreased productivity - No incentive for efficiency - Risk of lower quality (quality skimping) - No incentive to pay attention to the patient's demands - Provides little insight into practice profiles
Case payment	- Fairly good adjustment for case complexity - Fairly close relationship between remuneration and services - Provides physician with incentive to minimise the costs per case treated - Provides some insight into practice profiles - Facilitates integrated care	- Difficult to draw up an extensive list of diseases - The severity of cases (and the resulting costs) may vary considerably, even within a diagnosis group - Risk of patient selection by physician in case the severity of cases varies considerably within a diagnosis group - Risk of lower quality (quality skimping) - Risk of 'diagnosis creep' - Difficult to set a budget in advance

Source: Glaser (1970), Reinhardt (1985), Janssen (1988), Delnoy et al. (1992), and NRV (1993).

Ancillary payments
Besides using basic payment methods for primary care services, the third party may use ancillary payments as an additional incentive system. Such payments may be made to stimulate the provision of specific care (i.e. function-related fees, like a fee for a special kind of minor surgery or immunisation). Another possibility is to use such payments in order to stimulate the provision of cost-effective and high-quality care by the GP, or to stimulate him to arrange for the provision of such care by other providers (i.e. behaviour- or outcome-related fees). A behaviour- or outcome-related fee is an amount of money paid to the physician supplementary to the basic payment once he has, for instance, met a certain norm.

The difference between function-related, behaviour-related and outcome-related fees may be subtle but – as defined here – lies in two characteristics. A first characteristic is the presence of a check after a certain period of time or after a certain event (for instance, a treatment). As a result, the regularity of the payments may differ. The second characteristic is the nature of the check. The subject of the monitoring may be the behaviour of the physician, or it may be the effect of this behaviour.

Although a *function-related fee* is an ancillary payment, it will often have a continuous character. The payment may be made ex ante (or even ex post), but there is no check ex post. The fee may serve as a compensation for practice costs as well as remuneration for the physician himself. Hence the physician is given the opportunity, and thus stimulated, to provide certain services.

Contrary to a function-related fee, a *behaviour-related fee* is made contingent upon an ex-post check. There is the question of whether or how the physician provides the specified services before the definitive payment is made. Hence it is considered here to be an irregular payment, although it may become regular if the physician keeps succeeding in providing the services in a satisfactory way.

An *outcome-related fee* is also contingent upon an ex-post check. The question, however, is not whether or how the physician provided the specified services, but what the effects of his behaviour are. The effects may, for instance, be stated in terms of costs or volumes of care or in terms of health effects. Again, the fee is considered here to be an irregular payment, but it may become a regular payment if the physician keeps succeeding in meeting the required norm. Instead of just two possibilities (only a behaviour-related fee if the physician provided the specific services or if he provided them well) there may be a whole range of attainable qualities. Depending on what is agreed upon, all those qualities below or above a specified level may result in a bonus. Although not necessarily so, the outcome-related payment may vary to the extent that the norm is met.

The difference between the three types of ancillary fees may be illustrated by an example regarding an electronic prescription system. A third party may pay:
- a function-related fee in order to enable as well as to stimulate the physician to use such a system;
- a behaviour-related fee if the physician indeed has used the system correctly in the past year;

- an outcome-related fee if the system resulted in a lower number of prescriptions or in lower drug costs.

Another example is a payment for offering immunisation services, where the outcome-related payment is made once the physician has increased the immunisation rate of the eligible population beyond a certain level.

The category of ancillary payments consists of bonus systems, withhold systems and budget systems. In a *bonus system*, a GP receives an additional payment in case he meets a certain norm (financial or non-financial). For instance, he may receive a bonus if his actual costs are lower than the costs projected by the third party. Other examples are predetermined levels of patient satisfaction, medical knowledge and technical skills, productivity, compliance with pre-certification and authorisation requirements, co-operation with quality-management programs, use of network providers, or all kinds of utilisation figures (see also Kongstvedt 1993c, pp. 192-197). Besides on individual behaviour, a bonus may also depend upon total plan performance (Kongstvedt 1993a). The opposite of a bonus is a malus. In case of a malus system the physician has to return an amount of money for not meeting a certain (financial or non-financial norm), for instance if his actual costs are higher than the costs projected by the third party.

In case of *withholds*, a portion of the basic payment is withheld until a certain point in time (for instance, the end of the year). At that point, a physician's actual costs are compared with the projected costs, as in a bonus system. If the actual costs exceed the projected costs, the withhold is used to cover the excess costs. Any remaining money is paid to the physician (Kongstvedt 1993a).

Often, malus and withhold systems are considered to be incentive systems, but they may also be viewed as some kind of directive. The main difference between incentives and directives is that in case of incentives the physician is rewarded if he chooses those actions preferred by the third party. If he does not, he misses his reward. In case of directives, the physician is punished if he does not choose the dictated actions. So contrary to incentive systems, directives result in a declining welfare if the physician acts in another way than preferred. A physician can not claim the withheld sum if it appears at the end of the year that he did not practice efficiently.[10]

Bonus systems and withholds are frequently used to stimulate GPs to behave efficiently with regard to follow-up care, that is to order, refer and prescribe in an efficient way. GPs may also receive a *budget* for the costs associated with this care. An example is the British General Practice Fundholding scheme, in which GPs received a budget with which they had to finance health care services, like hospital care, diagnostic tests, drugs and appliances (see section 7.3).

The third party may blend basic payment systems to mitigate the negative effects and to combine the positive effects of the different systems. The combination of capitation and fee for service, for instance, may balance the incentives for under- and over-provision of

[10] Another way to look at this matter is to consider a directive to be a negative incentive (see also subsection 3.3.3).

care. In combination with the ancillary payments, this results in a number of payment options. An interesting option for paying the GP then may be a combination of a capitation payment and a bonus system for efficient referral behaviour. There are several arguments for blended or mixed payment systems. One argument is that mixed systems reduce the incentives for undesirable behaviour, like cream skimming and quality skimping, while maintaining some incentives for efficient behaviour (Ellis and McGuire 1990, Newhouse 1996). Another argument is found in case the third party wants to link the payment system to the performance of the physician in providing quality care. Since some dimensions of quality will never be contractible, mixed payment systems balance incentives for quality efforts across contractible and noncontractible dimensions of quality. This refers to the problem of multitasking: the difficulty of designing an incentive system to stimulate certain behaviour or a certain outcome if the provider has to perform various tasks and if for some tasks the desired outcomes are more difficult to measure (Eggleston 2005).

Application of payment methods

Hillman et al. (1992) surveyed contractual arrangements between IPA-type HMOs and primary care physicians. They distinguished between two- and three-tiered HMOs (see subsection 2.3.2) and found that 41 percent of the two-tiered HMOs used a fee-for-service system to pay primary care physicians, 45 percent used a capitation system, and 14 percent used a salary system. Twenty-six percent of the three-tiered HMOs used fee for service as final payment system for primary care, 18 percent used a capitation system, and 39 percent used a salary system.

In a survey among physicians, Remler et al. (1997) found that 43 percent of the individual generalists was paid a salary. Further, generalist practices received capitation payments for, on average, 18 percent of their patients. However, individual generalists received a capitation payment less frequently: on average for 9 percent of their patients. Apparently, part of the practices translated the capitation into another payment system. Of the generalists receiving some capitation (31 percent), the share of patients for whom capitation was paid, was 29 percent.

Gold et al. (1995a) found the following figures. Group- or staff-model HMOs paid their primary care physicians according to a salary system (28 percent) or a fee-for-service system (3 percent), or used other arrangements of which capitation (34 percent) was the predominant method.[11] Of the network- or IPA-type HMOs, 2 percent paid a salary, 12 percent used fee for service, and 84 percent used other arrangements with capitation (56 percent) as the predominant method. PPOs used fee for service (90 percent) or other arrangements with capitation (7 percent) as predominant method. Often, managed care-organisations adjusted the payments for utilisation measures. Fifty percent of the group- or staff-model HMOs made such adjustments; 74 percent of the network- or IPA-type HMOs did so. Of the PPOs, 34 percent made these adjustments too. Other payment adjusters were patients' complaints, quality measures, consumer surveys, physician pro-

[11] The category 'other arrangements' consisted of risk-sharing arrangements, like capitation with or without withholding or bonuses, or salary or fee for service with withholding or bonuses.

ductivity, or enrolee turnover rates.

In a more recent study among 116 HMOs, Gold et al. (2002) found that 24.7% of the HMOs paid their individual primary care physicians according to a fee-for-service system. About 8% combined this with a withhold or bonus system. About 61% paid a capitation, with about 29% combining it with a withhold or bonus. Around 14% paid a salary, with 3% combining salary and withhold or bonus. Interesting is the amount of care covered by the capitation fee:

- all HMOs included primary care office units in the capitation;
- 93% included other services provided in the physician office;
- 84% included ambulatory care provided elsewhere;
- 84% included inpatient visits;
- 47% included referral for specialist care;
- 46% included ancillary care provided by others.

According to Gold et al., the latter two are more likely to be included when payments are made to a group of physicians instead of to individual physicians.

Hemenway et al. (1990) described the effect of implementing a bonus system for 15 physicians in an ambulatory care centre in the US. The implementation resulted in an increase of the number of services. Patient visits rose by 12% and charged services rose by 20% per month.

Flierman (1991) and Delnoy et al. (1992) described the effects of a change in payment system for the GPs in Copenhagen. The capitation system was replaced by a blended system of capitation and fee for service. Both payment systems generated 50 percent of the physician's income. The introduction of the blended payment system resulted in a significant increase in the number of diagnostic and curative services in the experimental group compared to the control group.[12] For instance, the number of diagnostic services showed an increase of 66% in the experimental group versus an increase of 11% in the control group. The number of curative services showed an increase of 80% in the experimental group versus an increase of 18% in the control group. The number of referrals to private medical specialists or to hospitals decreased with respectively 27% and 23% in the experimental group versus respectively 1% and 6% in the control group.

Scott and Hall (1995) reviewed 18 studies of the effects of several remunerating methods for GPs/primary care physicians. They concluded that it is difficult to evaluate these effects due to methodological problems. Randomised-controlled trials are seldom feasible. Further, it requires changing payment systems or the presence of different methods within the same setting. According to Hellinger, many studies are subject to several sources of bias, like patient selection effects, physician selection effects and missing variables (Hellinger 1996).

Sørensen and Grytten (2003) found that Norwegian primary care physicians who were paid on a fee-for-service basis had more consultations, more patient contacts and less referrals than their colleagues who were paid on a salary basis. The differences were

[12] The research was not a randomized-controlled trial, but the group of physicians confronted with the change of the payment system (physicians in the city of Copenhagen) was compared with a control group (physicians in the province of Copenhagen).

mainly the result of a higher number of working hours, but also of higher time efficiency. These results were for about one-third due to a selection effect and for about two-thirds due to the incentive effect. In the long-term, however, the selection effect seemed to fade away, which may be due to altered physician preferences. Sørensen and Grytten expected that changing the salary system into a fee-for-service system would result in an increase of these physicians' their production by about 23% in the short-term and by about 40% in the long-term.

As an exception, Hickson et al. (1987) described the results of a randomised-controlled trial in which the impact of payment systems on the behaviour of the participating (18) physicians was tested. They found that physicians who were reimbursed on a fee-for-service basis scheduled more visits per patient than salaried physicians did and saw their patients more often during the study of nine months. This difference was mainly the result of fee-for-service physicians seeing more well patients than salaried physicians.

In chapter 7 we will discuss other studies of the effects of financial incentives on the behaviour of physicians.

5.4.1.2 Non-financial incentives

Gatekeeping

In general, a gatekeeper model is an arrangement in which a GP must give prior authorisation for diagnostic tests and several forms of care, like community health services, non-emergency specialist care, hospital care, and sometimes also emergency care. Payment for these services is often linked to the referral by a primary-care gatekeeper (Trapnell 1985, Franks et al. 1992). Nevertheless, gatekeeper models may differ in the degree to which they serve as a prior-authorisation system, and to which the gatekeepers function as organisers and co-ordinators of care (Gold et al. 1995b).

Primary-care gatekeeping may in the first place be viewed as a technique to direct patients' behaviour, for in a gatekeeper system patients are required to have the gatekeeper authorise specialist care, hospital care, et cetera. Yet primary-care gatekeeping is also a technique to control the GPs themselves. One could consider the appointment of GPs as gatekeepers to be a directive. Here, however, assigning them the gatekeeper function and providing them with the necessary tools is viewed as a stimulus to actually perform this function. Ultimately, it will be these physicians who decide whether or not to refer patients (thus whether or not to respond to the incentive). The individual autonomy of the physicians as well as the professional autonomy of the profession is thus maintained. As incentive, the appointment and provision of tools may not be strong enough and physicians may still be triggered to order, refer and prescribe in a sub-optimal way. Therefore, gatekeeping often coincides with other techniques, like financial incentives.

As with other managed-care techniques, primary-care gatekeeping may be considered a cost-containment device or a technique to improve the quality of care. It is a way to control costs as the number of visits to other and perhaps more expensive providers of care, like specialists or hospitals, may be reduced (the inhibiting effect of gatekeeping). In addition, it may result in a shift from more expensive secondary care towards less ex-

pensive primary care (the substitution effect). Franks et al. (1992) argued that viewing gatekeeping as a function of opening and closing the gate to costly care is a simplistic view, because it suggests that the GP is the third-party payer's agent instead of the patient's. They defined gatekeeping as 'the process of matching patients' needs and preferences with the judicious use of medical services' (Franks et al. 1992, p. 424). It should be noticed that since in this thesis the third party is supposed to act as agent on behalf of the patients, a GP who acts as agent for the third party should, ideally, act identically as a GP who acts as the patient's agent. If primary-care gatekeepers indeed have to match their patients' needs and preferences with the use of health care and also have to organise and co-ordinate care, then they have adopted a large part of the third-party's agency function. In fact, part of the decision-making process has been delegated to the GP then. Although one may argue that such decisions should have been the GP's in the first place, not seldom this decision-making has become part of the third party's domain. This is, for instance, exemplified by the use of techniques like pre-admission review and mandatory second opinions (see also subsection 5.4.3).

Gatekeeping may also improve the quality of care. Although undertreatment is a serious risk in gatekeeping arrangements and may require strategies to prevent this risk, protecting patients from overtreatment is a major benefit. Gatekeeping may reduce the number of false positive test results and iatrogenic diseases (Franks et al. 1992). It may also contribute to the reduction of medicalisation. This illustrates that quality improvement might very well result in cost containment and vice versa. But primary care gatekeeping has some other advantages. As the patient has to visit his GP each time he might need medical care, the physician gets to know him better and is able to form a notion of the patient's medical history and to get a picture of his life and work. This also enables the physician to keep a medical record. Another advantage is that the gatekeeper may support the patient in making decisions on the required care and the most suitable type of provider. Furthermore, for the patient treatment by a GP may be more pleasant than if he has to visit a specialist or a hospital. A final advantage mentioned here is that the gatekeeper system enables medical specialists to maintain and to develop their skills and knowledge, as the group of patients they see is much more a selected group than it would be absent the filter of the gatekeeper.

Remler et al. (1997) found that generalists act as gatekeepers for, on average, forty percent of their patients. Within 61 percent of the group- or staff-model HMOs patients are required to select an individual primary care physician, and 96 percent of these organisations holds their primary care physicians responsible for most specialist referrals. Ninety-two percent of the network or IPA HMOs required their patients to select an individual primary care physician who had also responsibility for most specialist referrals (Gold et al. 1995a).
Also in the Netherlands and the UK GPs act as primary care gatekeepers for referral care.

Group-membership incentives
Physicians may be working in a group practice or may be member of a risk pool. [13] Apart from the direct financial incentives emanating from membership of a risk pool, there are *group-membership incentives* as well. The proximity of peers, the membership of a risk pool and the pressure to comply with group norms or standards may influence provider behaviour and, therefore, can be regarded as another technique the third party may use. Eisenberg (2002) argued that in every practice organisation there are physicians who are particularly influential in determining the group norms of practice style. The role of clinical leadership by educationally influential physicians can be used as a starting point to influence the practice patterns of their colleagues.

5.4.2 Control by persuasion/information

A third-party agent may inform a GP about the desired actions hoping that the physician values these actions more positively then, and may try to persuade him to perform these actions. The techniques discussed here briefly are practice guidelines (and clinical rules), physician profiling and high-cost case management.

Practice guidelines
In the course of time, various names have evolved that more or less embody the same concept, i.e. instructing a physician how he should or should not act in a specified situation. Examples are practice guidelines, protocols, clinical rules, regulations, and prospective utilisation review (Hillman 1991). As different authors use the terms differently, it is not always instantly clear which exact meaning their terms have.

Practice guidelines, protocols or clinical rules may be developed for diagnostic as well as therapeutic services. Although the terms more or less embody the same concept, in practice they may differ in the way the physician is obliged to follow the instructions and is sanctioned if he does not do so. Here, practice guidelines are considered to have an advisory role; the physician will not be punished by the third party if he acts differently. The third party may thus employ practice guidelines to inform the physician and to persuade him to perform certain actions, i.e. to make the physician value the preferred actions more positively. However, if the instructions are formulated in a coercive manner, as protocols or rules, then they may be classified as a directive. Protocols and clinical rules are thus considered to be compulsory by nature (see subsection 5.4.3). An example of practice guidelines are the eighty NHG-Practice Guidelines, developed by the Dutch College of General Practitioners (NHG).

All these techniques may coincide with, for instance, financial incentives. In case of practice guidelines, a physician may be rewarded financially in case he uses them. In case of protocols and clinical rules, a physician may be confronted with withheld payments or even with termination from a contract.

[13] A risk pool is a group of providers, like GPs, who share in the rewards or penalties from surpluses or deficits in the budgets for certain types of care, for instance secondary care (Hillman et al. 1992). To this subject is returned in the next chapter.

Practice guidelines are usually based on codified approaches to health care services. Nevertheless, physicians often judge them negatively, as they are associated with cook-book medicine (Kongstvedt 1993b). Hillman (1991, p. 139), who considers 'clinical rules' to be the collective noun that also comprises practice guidelines, mentioned several advantages of clinical rules (and thus of practice guidelines). Firstly, they are a way to inform physicians about appropriate and effective treatments. Secondly, they facilitate the evaluation of compliance with new clinical approaches and they may support physicians if they want to withhold a service a patient demands for. Thirdly, rules provide a basis for standardising the approach to a certain problem. Fourthly, rules may support physicians in defending themselves against malpractice claims if they can prove compliance with a rule issued by the profession.

Remler et al. (1997) found that generalists were subject to condition specific guidelines or protocols for, on average, 17 percent of their patients. Gold et al. (1995a) found that 76 percent of the group and staff HMOs as well as of the network or IPA HMOs made any use of practice guidelines. Twenty-eight percent of the PPOs did too. Any use of guidelines involved establishing formal, written guidelines, monitoring compliance, and meeting with physicians to review results.

Physician profiling

By means of physician profiles or practice profiles, GPs may be informed about their relative performance. Profiling entails '(...) the preparation and selective dissemination of reports that compare the practice patterns of different providers on such dimensions as resource consumption, charges, and outcomes. (It) provides relative performance measures among providers, and can be used to identify potential quality problems, assess provider performance, and improve utilisation review' (Evans et al. 1995, p. 1107). As the total care provided by a physician to his practice population is reviewed, the focus is on practice patterns rather than on the uniqueness of one case (Parkerton et al. 2003). Profiling serves as a feedback mechanism.[14]

High-cost case management/large-case management/medical case management/ catastrophic-case management/individual benefits management

'High-cost case management' refers to the process of identifying patients with (expected) high costs, assessing their needs and personal circumstances, and then arranging less expensive care, preferably of at least the same quality. It thus differs from the other utilisation-management techniques in that it is specifically aimed at (potential) high-expenditure patients and at arranging alternative care as well as assessing the appropriateness of the care.

A case-management program is a voluntary technique (IOM 1989). It is considered here to be employed by a third party in co-operation with the GP. The physician, however, remains responsible for the final health-care decisions. Frequently, specialised

[14] Besides being a means to provide physicians with feedback, profiling is also a monitoring technique used in order to provide the third party with information about the monitored physicians. For a further discussion of profiling, see section 5.5.

nurses provide case management. In that instance it is no longer a means to support the GP in making health-care decisions but an autonomous treatment by another provider of care. It is beyond the scope of this research then.

5.4.3 Control by directive/authority

A third method the third-party agent may use in an attempt to control a GP is by means of directives or authority. The actions the physician has to perform are defined, so his choice is restricted. Of course, even being subjected to directives the physician may still choose to perform otherwise but he risks sanctions then. Several techniques are used in health care, all of which can be ranged under the term directive but are often summarised by the term 'utilisation management'. A selection of these techniques is explained here.[15]

Utilisation-management techniques form a well-known part of managed care. Less well known, however, is the precise meaning of the term. Several definitions circulate and authors range different techniques under the term. One was proposed by The Committee on Utilisation Management by Third Parties of the Institute of Medicine (IOM). The committee defined utilisation management as 'a set of techniques used by or on behalf of purchasers of health care benefits to manage health care costs by influencing patient care decision-making through case-by-case assessments of the appropriateness of care prior to its provision' (IOM 1989, p. 17). For four reasons, this can be considered a narrow definition. Firstly, the committee mentioned solely *purchasers* of health care and not, for instance, group practices. Secondly, according to the committee, goal of utilisation management is to manage *health care costs*. The reduction of unnecessary care is left out of consideration. A third reason is that only *case-by-case assessments* of the appropriateness of the care are taken into account. Finally, the committee focused on techniques used to assess the care *prior* to its provision. The committee distinguished two main ex ante techniques: prior review and high-cost case management. Prior review techniques were further subdivided into pre-admission review, admission review, continued-stay review, discharge planning and second opinion.

Bailit and Sennett (1991, p. 87) broadened the IOM definition by omitting the last part of it, 'prior to its provision'. As a result, utilisation management also includes retrospective utilisation review. Another definition was proposed by Milstein (1997, p. 87), who described utilisation management as 'all interventions originating outside the physician/patient relationship with an intent to promote an economical mix of health care services'. Kerr et al. (1995) defined utilisation management in the broadest sense; it included physician incentives and primary care gatekeeping.

Striking is the emphasis on the costs of care instead of the quality of care; just as is the case with many definitions of managed care (see subsection 5.2.1). In a survey among utilisation-review organisations, respondents reported that in one to three percent of the cases they reviewed in 1993 they observed 'unnecessary or inappropriate care that in-

[15] The description of the terms is mainly based upon the study of the Committee on Utilisation Management by Third Parties of the Institute of Medicine (1989). See also Langwell (1990) and Gold et al. (1995a,b).

volved significant risks for patients' (Schlesinger et al. 1997, p. 116). Although this percentage may seem low, it indicates that maintenance of and improvements in the quality of patient care are attainable goals. Moreover, Schlesinger et al. extrapolated this figure to the whole industry and estimated that utilisation-review organisations identify between 85,000 and 255,000 of such cases per year. Further, utilisation management may contribute to the quality of care in the other 97 to 99 percent of the cases, although in a less dramatic way.

Whether the techniques within the second category, the control techniques, are considered to be stimulating, persuading, or directing, will mainly depend upon the sanctions. In case of utilisation management, a sanction may be that the third party does not pay for the services if the physician ignores its treatment proposals. The physician then faces a loss if he perseveres despite the reviewer's directives.[16] Utilisation management should then be considered a directive. Case management is an exception to this as, usually, case-management programs are voluntary (IOM 1989). It may be considered an incentive then, or a way to persuade physician and patient to use the recommended services.

Clinical rules, protocols

A first method to direct the behaviour of GPs is to issue clinical rules, or protocols, stating how a physician should or should not act, given a certain clinical circumstance. Imposing clinical rules can be considered a first step as it is taken before a patient comes in view. See also subsection 5.4.2.

Pre-admission review/pre-admission certification/pre-service review/ pre-procedure review/prior authorisation

Although the terms may be mixed up, generally 'pre-admission review' is used to indicate a technique that is employed to assess the need for a proposed hospital admission, prior to the admission itself. 'Pre-admission certification' indicates that an admission needs to be certified in order to obtain payment. 'Pre-service review', 'pre-procedure review' and 'prior authorisation' are usually used to indicate the assessment of the need for a procedure, regardless of whether it will be performed on an inpatient basis or not.

The difference with gatekeeping is that the current techniques are applied by a third party. The appointment of the GP to the gatekeeper function may be done by a third party, but the performance of the function is done by the physician. Thus in case of pre-admission review, pre-admission certification et cetera, the third party is the final decision-maker, whereas it is the physician in case of primary-care gatekeeping.

Admission review

The term 'admission review' is used to describe a technique employed to assess whether an emergency or an urgent admission was appropriate or not. In case of such admissions,

[16] It may make a slight difference whether a patient is reimbursed or whether the services are delivered in kind: in a reimbursement system it will be the patient who bears the costs of the reviewer's sanctions.

pre-admission review is generally not feasible so that the admission is reviewed within a few days after hospitalisation.

(Mandatory) second (surgical) opinion
For some procedures patients have to get a second opinion from another physician to check whether the proposed services are needed and appropriate.

Continued-stay review/concurrent utilisation review/length-of-stay review
These are techniques used to check whether continued inpatient care is really needed or whether a patient could also be treated on an outpatient basis. The term 'concurrent utilisation review' may also apply in case the third party assesses the appropriateness of the hospital care itself. The inpatient basis does not have to be under discussion then.

Discharge planning
'Discharge planning' refers to techniques used to ensure that patients are discharged from hospital as soon as medically justified. These techniques may range from indicating, at admission, an expectation of the length of stay to extensive post-discharge service planning.

Retrospective utilisation review
If is assessed whether the already delivered health services were appropriate, then is spoken of 'retrospective utilisation review'. Although this is a technique that is applied after care has been consumed, to a patient the threat of denial of claims may be an incentive strong enough to refrain from over-consumption. To a GP the threat of termination of his contract by the third-party agent is an incentive to refrain from demand inducement. Except for using it to assess the care delivered to a specific patient, it may also be employed:
- to monitor whether the information provided prior and during the delivery of services was accurate;
- to examine the high-volume, low unit-cost claims, which are not suitable for assessment prior or during service delivery;
- to analyse patterns of care for physician, practice or hospital profiling and for selective-contracting purposes (IOM 1989, p. 20).[17]

Application of utilisation management
Of all managed-care techniques, utilisation-management techniques seem to be the most employed.[18] In a survey of U.S. physicians conducted in 1995, Remler et al. (1997) found

[17] It should be noticed that here is departed from the principle of case-by-case assessment as mentioned by The Committee on Utilisation Management by Third Parties of the Institute of Medicine (IOM).
[18] Since all third parties use some payment method and since such methods all provide physicians with some kind of financial incentive, it may be argued that financial incentives are the most widely employed technique. However, it is questionable whether all third parties consciously choose a basic method of payment as a way to influence medical practice. Moreover, at least in some countries the payment method results from legislation. This may be different for risk-sharing arrangements, but these are less widely employed then utilisation-management techniques.

that on average 59 percent of a physician's patients had a health-insurance plan reviewing the length of hospital stays. For surgeons, this number was even somewhat larger: on average 62 percent. On average 58 percent of a generalist's patients had their length of stay reviewed. On average 45 percent of a generalist's patients had their site of care reviewed and 38 percent the appropriateness of diagnosis and treatment.

Gold et al. (1995a) found that of all managed-care plans they surveyed in 1994, 62 percent used at least four of the following five utilisation-review techniques: pre-admission review for all non-emergency admissions, concurrent and retrospective review, discharge planning, and ambulatory review for resource-intensive services. Ninety-five percent used one of these techniques.

The reviewing party may deny coverage for services a physician recommended his patient. Remler et al. (1997) found that for the several forms of care surveyed, the maximum percentage of denials was six percent of the patients for whom care was recommended. Second-round denials (as a result of successful appeals) were at most three percent. For example, first-round denial rate of hospitalisations was 3.4 percent; second-round rate was 1.0 percent. Generalists were less likely than medical specialists to have coverage denied for hospitalisations, referrals to specialists, endoscopies and cardiac catheterisations, but were more likely to experience a denial for substance abuse referrals, mental health referrals and MRIs. Striking are the differences between physicians: most physicians did not experience a denial at all, whereas five percent of the physicians experienced a denial for at least a fifth of their mental health referrals. One percent had a denial for at least twenty percent of the hospitalisations they recommended.

The rates of first-round and second-round denials by utilisation review organisations found by Schlesinger et al. (1997) ranged from on average 3.6 percent and 2.8 percent respectively to 14.1 percent and 12.5 percent respectively of requested hospitalisations.[19]

Remler et al. (1997) noted that although the proportion of final denials was always at most three percent, utilisation review may have a larger impact since its presence may discourage physicians from recommending care for which it may be expected that the reviewing organisation will deny coverage. Moreover, both Remler et al. and Schlesinger et al. did not present figures of conversions or withdrawals. Conversions are those cases in which physicians are convinced by the reviewer to change a recommendation for hospitalisation in outpatient care, and where coverage is not formally denied. Withdrawals

[19] At least two reasons may explain the differences in denial rates between the surveys of Remler et al. (1997) and Schlesinger et al. (1997). The first reason is that the percentages found by Remler et al. reflect the share of patients for whom care was recommended and coverage was denied. The patients for whom care was recommended, however, included patients for whom the care was not reviewed by their health insurance plan. So, the share of patients for whom care was recommended and for whom coverage was denied, provided that their care was reviewed, will be higher. The percentages found by Schlesinger et al. reflect the share of patients for whom the care was reviewed and not covered. A second reason for the differences in denial rates may be that the reviewing organisations in the Remler survey are health insurance plans; in the Schlesinger survey the care is reviewed by utilisation review organisations. Utilisation review organisations are specialised organisations. Moreover, they are hired by other third parties to review their members' care and it may, therefore, be hypothesised that they use utilisation-management techniques more aggressively in order to come up to the expectations.

are cases in which a recommendation for hospitalisation is dropped before the reviewer formally denies it (Schlesinger et al. 1997).

5.5 Monitoring general practitioners

Physician profiling

Physician profiles or practice profiles may be used to monitor the behaviour of GPs and to analyse the outcomes.[20] Results can be compared with results from other providers or with guidelines. The method differs from many utilisation-management techniques in that it is retrospective instead of prospective or concurrent and that it does not focus on individual cases but on patterns of care (Welch et al. 1994).

Two main goals of profiling can be discerned. One goal is to inform the less-informed physician. A third party may be better informed about the relative performances of physicians. Providing a physician with information about his relative performance gives him an opportunity to review his own behaviour and to adjust it if desirable. Physician profiling may, therefore, be considered a technique to control the physician by means of persuasion or information. It is thus a way to reduce the information gap between third party and physician by informing the physician.

The other main goal of physician profiling is to monitor the physician. This may reveal information about the physician's behaviour as well as the outcome. As noted in section 4.2.2, the outcome only partially results from the physician's actions, and will result from the natural course of the disease or other health-influencing factors as well. Trying to monitor the physician's behaviour may reveal information about the appropriateness of the physician's actions and the quality of the physician as such. The third party may use this information then for selective contracting purposes. Besides being a way to reduce the information gap between third party and physician by informing the physician, physician profiling is thus a way to reduce the information gap between physician and third party by informing the third party.

A third party may monitor a physician and compare the findings with a practice-based norm or a standard-based norm (Evans et al. 1995). Practice-based norms originate from the experiences of comparable providers. These may be located within, for instance, the same practice, hospital, or region. Standard-based norms are derived from practice standards. The other way around, physician profiles may be used to specify practice standards for specific diagnoses or procedures (Boland 1985). Practice-based norms are more widely used than standard-based norms (Gold et al. 1995b).

Parkerton et al. (2003) argued that assessment of the performance of primary care physicians requires multiple, reliable measures. They pointed at the fact, though, that it is still unclear what the relationship is between the several measures of physician performance and between these measures and costs. The physician's performance with a certain type of disease is probably not a reliable indication of his performance with other types of

[20] See subsection 5.4.2 for a definition of profiling.

diseases. His performance with a specific type of screening may not provide a good indication of his screening in general.

Remler et al. (1997) found that the physicians surveyed were subject to profiling for on average 16 percent of their patients. For generalists this percentage was, on average, 22 percent. Gold et al. (1995a) found that 74 percent of the surveyed managed-care organisations used profiles, provided physician feedback, or identified areas for system-wide improvement. For group or staff HMOs, network or IPA HMOs, and PPOs the percentages were 76, 86, and 52, respectively.

Evans et al. (1995) analysed the effect of a hospital's introduction of a program to profile the length of stay of the physicians' patients.[21] The profiling program had a statistically significant effect on the share of physicians who achieved the practice-based length-of-stay benchmark. The reductions in the length of stay depended on:

- the physicians' initial performance level (with reductions mainly achieved by physicians who initially failed to meet the norm),
- patient severity (with improvements primarily for patients at medium severity levels),
- and the economic significance of the DRGs, the Diagnosis Related Groups (the larger the impact, the larger the effect).

5.6 The use of managed-care techniques

Third parties can apply managed-care techniques independently or complementary to other techniques. For instance, financial incentives as well as practice guidelines may be used on their own or in combination. In the latter case, payment of financial rewards may have been made contingent upon the use of practice guidelines as to stimulate its use.

Little is known about the exact arrangements third parties make with providers of care (Gold et al. 1995a,b, Remler et al. 1997, Glied 2000). Gold et al. (1995a) tried to identify the arrangements made between managed-care organisations and physicians. It appears from their survey that managed-care organisations, and in particular HMOs, have complicated systems to select, control and monitor physicians. Although their article provides a wealth of information about the arrangements, it has some limitations. Firstly, it offers little insight into the effects such arrangements have on the cost, quality and accessibility of the care (Gold et al. 1995a, p. 1682). Secondly, although it provides information on the extent managed-care organisations employ managed-care techniques, it does not provide information on the extent of the exposure of physicians to these techniques. This is an important limitation since physicians often have patients from several third parties, including non-managed indemnity plans (Remler et al. 1997). Thirdly, it would be interesting to know whether the techniques employed are substitutes or whether they are used complementary. Neither the Gold article nor the Remler article provides much information on that subject. Nevertheless, useful information on the application of managed-care techniques can be deduced from their research. The following table is compiled of findings by Gold et al. (1995a).

[21] These figures may be different for a profiling program introduced by a third-party agent.

Table 5.2 Techniques used by third parties within their relationships with physicians, US 1994 (figures in percentages; source: Gold et al. 1995a).

Technique	**All plans** (N = 108)	**Group or staff HMOs** (N = 29)	**Network or IPA HMOs** (N = 50)	**PPOs** (N = 29)
Board certification or eligibility requirement	57	90	48	41
Large emphasis on previous costs or utilisation patterns in selection decisions	13	4	18	14
License and credentials verification	100	100	100	100
Office visit, facility review, medical record screening: - all of these - none of these	43 27	38 34	66 8	7 52
Quantitative data review	37	24	38	48
Financial-risk sharing between third party and PCP: - capitation as predominant method Salary (no withholds or bonuses) Fee for service (no withholds or bonuses)	60 37 8 31	68 34 28 3	84 56 2 12	10 7 0 90
Payment adjusters: - utilisation or cost measures - patient complaints - quality measures - consumer surveys - provider productivity - enrolee turnover rate - none of these	 57 49 46 36 24 21 28	 50 57 54 37 43 11 29	 74 61 64 55 26 36 14	 34 21 7 3 3 3 55
PCPs responsible for referrals to most specialists	94	96	92	–
Quality monitoring and focused studies: clinically focused studies, outcome studies, quality improvement initiatives, and the use of these for quality improvement or success measurement: - all of these - focused studies regularly	 62 83	 79 100	 70 96	 31 45
Use profiles, provide physician feedback, search areas for improvement: - all of these - one of these	 68 74	 69 76	 80 86	 45 52
Use formal practice guidelines, use them extensively, monitor compliance, meet with physicians to discuss results: - all of these - one of these	 26 63	 31 76	 34 76	 7 28

Technique	All plans	Group or staff HMOs	Network or IPA HMOs	PPOs
Pre-admission review for all non-emergency admissions, concurrent and retrospective review, discharge planning, and ambulatory review for resource-intensive services:				
- at least four of five	62	72	70	37
- one of these	95	97	100	86

As stated, the figures do not give an exact insight into the way managed-care organisations combine the several techniques. However, several techniques are applied more often, so it is more likely that they are used complementary. For instance, it is likely that the majority of network or IPA HMOs that shared their financial risks with primary care physicians (84 percent) consists of the same HMOs that made their primary care physicians responsible for the referrals to most of the specialists (92 percent). It is hard to deduce, though, whether the PPOs that apply profiles (45 percent) are the same PPOs that use incentives based on performance by adjusting payments to primary care physicians (45 percent). Possibly, a proportion of the PPOs does use profiles but does not use them to adjust payments.[22] Without knowing exactly which (combinations of) techniques are applied within which settings, it is difficult to judge the effectiveness of the techniques.

Judgement of the effectiveness is further hampered by methodological problems, as there are virtually no randomised-controlled trials. Steiner and Robinson reported on a large review of the evidence on managed care in the US. Regarding the effectiveness of managed-care techniques they concluded the following: 'Despite literally thousands of publications since 1990 whose subject is some component of the managed care approach, it is still not possible to answer fundamental questions about the independent contribution of each component to organisational performance. There are almost no randomised-controlled trials of these techniques in managed care settings. Most publications either describe or advocate the use of techniques, without any evidential basis; many others evaluate interventions only in qualitative terms, lack comparison groups, and make no tests of statistical significance' (Steiner and Robinson 1998, p. 178).

Glied (2000) reviewed the managed-care literature extensively and argued that the empirical research on managed care is complicated by two factors. Firstly, managed-care plans use different combinations of managed-care techniques, or use the same techniques in a different way (for instance in a more stringent or in a less stringent way). Research also suffers from the heterogeneity of plans. For instance, some plans are for-profit, whereas other plans are not-for-profit, and some plans are insurer-based, whereas other plans are provider-based. Secondly, as a result of risk segmentation, managed-care enrolees may differ from enrolees of conventional insurance plans ('remote third-party pay-

[22] It may be argued that it is not likely that the PPOs that do not apply profiles do adjust payments to primary care physicians. However, it may be the case that some PPOs use other information than cost and utilisation patterns for payment adjustment. For instance, consumer satisfaction or the physician's use of practice guidelines may be used for this.

ers'). Hence differences in the use and outcome of care may be the result of characteristics of the enrolees instead of the use of managed-care techniques.

More information is available on the overall performance of organisations that apply managed-care techniques in any form. Schut (1986), for instance, reviewed a large number of studies related to HMOs. For preventive care he found that HMO-members had a higher use of preventive care than the traditionally insured, provided that the HMO-members had a better insurance coverage for this type of care. In case of comparable insurance coverage, the use of preventive care by HMO-members was equal or lower. The volume of hospital care was lower within HMOs than within the traditional insurance sector. One type of HMO, the Prepaid Group Practice (PGP) had 20 to 40% less hospital admissions than the traditional system. The number of hospital days was about 35% lower, but the length-of-stay was comparable to the traditional system. Remarkably, the use of ambulatory care was also comparable to the traditional system. The costs of care for PGP members were 10 to 40% lower than the costs of care for the traditionally insured. Lower costs mainly resulted from the lower volumes of (hospital) care. According to Schut, the quality of care was at least comparable to the quality provided within the traditional system. For some specific items, like waiting times and the patient-physician relationship, patient satisfaction seemed to be lower though.

Also Miller and Luft (1994) reviewed a large number of studies (including the randomised-controlled RAND health insurance experiment) that compared managed-care organisations with non-managed indemnity plans on the following items: health care utilisation, expenditure, prevention, quality of care, and enrolee satisfaction.

Health care utilisation

With regard to *hospital admission rates*, they found generally lower rates for HMOs. The differences between HMOs and indemnity plans varied considerably per study, though, ranging from no statistically significant rates, to 26 to 37 percent fewer hospitalisations. There were no significant differences between the several types of HMOs (staff, prepaid group practice, network, or IPA). The *length-of-stay in hospitals* was generally 1 to 20 percent shorter for HMOs. No major differences were found between the different HMO types. The *physician office visits per enrolee* were balanced for older studies in which the number of observations with lower use within the HMO setting equalled the number of observations with higher use. The number of visits did differ in studies with more recent data, indicating the same or more physician office visits for HMO enrolees. However, no evidence was found of higher use of physician services for HMOs as to compensate the lower hospital use. Further, on average HMOs used 22 percent fewer *services that are expensive and/or have less costly alternatives.*

Expenditure

Only a few studies were found in which the effects of managed care on costs were studied. In two studies the total expenditures per enrolee were lower within HMO settings. In one of these, 13 percent lower expenditures were found. In the other the difference was 11 percent. Only the first study showed a significant difference.

Prevention
Clearly, HMO enrolees were provided with more preventive tests, procedures and examinations. They also received more health promotion activities.

Quality of care
For the quality of care the results were somewhat ambiguous. Most studies showed better or equivalent process or outcome quality. However, some studies indicated that the quality was less for HMO enrolees. Two studies, for instance, showed adverse results for HMO enrolees with mental health problems.

Enrolee satisfaction
Satisfaction of enrolees can be split into satisfaction with costs and satisfaction with services. HMO enrolees were found to be more satisfied with the costs of their health plan than were enrolees in a non-managed indemnity plan. Regarding services, the studies showed mixed results, but HMO enrolees tended to be less satisfied. This was mainly because of restrictions they faced on the choice of physicians.

Although the results were not unambiguously favourable to HMOs, Miller and Luft suggested that HMOs provided services at lower costs than fee-for-service indemnity plans. This was based on the fact that HMOs used less inpatient services as well as services that are expensive and/or have less costly alternatives, and on the fact that HMOs provided enrolees more comprehensive coverage. Another major finding of their review was that, contrary to earlier findings, there was no difference in performance between the several HMO types. Prepaid groups or staff HMOs did not perform better than IPA or network HMOs. Miller and Luft (1994, pp.1517-18) have several possible explanations for this. Firstly, it may be due to methodological problems. The number of observations for the several HMO types was small. Secondly, physician groups in network HMOs may have started to employ financial incentives, like risk sharing, too. Thirdly, it may be easier to change medical practice within the newer primarily capitated groups than within the longer established prepaid group practices or staff HMOs. Finally, IPAs increasingly rely on utilisation management and financial incentives. Not only do these types of HMOs rely more on these techniques than in the past; they also do more than prepaid groups or staff-model HMOs.

In an update of their previous literature analysis, Miller and Luft (2002) found roughly comparable results. Compared with non-HMOs, HMOs provided more or less comparable quality of care, scored less on access to care, had lower ratings for enrolee satisfaction, provided more preventive care, showed shorter lengths-of-stays and used less expensive resources.

The findings of Gold et al. (1995a), as reproduced in the table, are consistent with the judgement of Miller and Luft that network and IPA HMOs rely more on financial incentives and utilisation-management techniques than do prepaid group practices and staff HMOs. Apparently, the IPA and network HMOs have caught up with the group and staff model HMOs. It may be the case that IPAs use more formal, contractual arrangements (i.e. individual pressure), whereas prepaid group practices use more informal arrangements and rely more on group norms (i.e. social pressure).

5.7 Summary and discussion

In order to find an answer to the research question central to this chapter,

Which techniques can and do third-party agents apply within their relationships with general practitioners in order to reduce the agency problems within the patient-physician relationship?

We gave an overview of techniques that are used by third parties rather commonly – that is to say, in some health-care systems. In health care (the use of) such a set of techniques is usually designated as 'managed care'. We argued that the managed-care techniques fit in the triptych of agency theory remarkably well. The three successive phases that comprise this triptych (selecting, controlling and monitoring the agent) form an iterative process, which we labelled the managed-care cycle. Hence from the perspective of agency theory, managed care can be viewed as (the cyclical use of) a set of techniques by which the third-party agent may attempt to influence the behaviour of the agent in a way that is beneficial to the patient.

The selection and contracting of (primary care) physicians is regarded to be crucial in managed care, in spite of the difficulties associated with it. Selection of physicians with a conservative practice style is considered to be the best guarantee of cost-effective and high-quality care. Research indicates that third parties prefer physician selection before concluding contracts above 'pruning later'. 'Pruning' may be easier, but it will be difficult to change the behaviour of physicians by then or to get rid of them afterwards. Selection and contracting of the selected physicians requires sufficient and reliable qualitative information and quantitative data about the physicians' behaviour controlled for case mix and practice size. Self-selection of physicians can reduce the selection problem though. Contracting the selected physicians requires an oversupply, legal possibilities to contract selectively, the co-operation of individual physicians or of the profession as a whole and the insured consent.

The second phase of the agency cycle is found in managed care as well. A prominent controlling means is the use of incentives, which may stimulate the physician to choose those alternatives from a set of possible actions that are most beneficial to the third party and to the patient. Financial incentives may emanate from the basic payment system and from ancillary payment systems. Ancillary payments may be function-related, behaviour-related or outcome-related. Contrary to a function-related fee, behaviour- and outcome-related fees are made contingent upon an ex-post check on the way the specified services are provided and on the effects of the physician's behaviour respectively. The category of ancillary payments consists of bonus, withhold and budget systems. The third party may use mixed (blended) payment systems in order to balance the incentives from the several basic and ancillary payment systems. A mixed system may, for instance, balance the incentives for undesirable behaviour, like cream skimming and quality skimping, and efficient behaviour as well as the incentives for quality efforts across contractible and non-contractible dimensions of quality.

The GPs' assignment to the gatekeeper function is another example of an incentive system. The physician is stimulated to refer patients to other providers of care only if necessary. As it is a weak incentive, gatekeeping is often combined with other techniques.

Other prominent means to control physicians are practice guidelines, physician or practice profiling and utilisation-management techniques. Some, like guidelines and profiling, are aimed at informing the physician and persuading him to perform the desired actions. Other, like (pre-)admission review, mandatory second opinion and continued-stay review, are aimed at restricting the choice of the physician. Some view these techniques as an infringement on the professional autonomy or the physicians' autonomy. However, infringement on the individual physicians' autonomy is exactly what is considered as a way to ensure that the patients' interests are served best. It depends on the role of the profession in designing and conducting managed care whether the techniques are an infringement on the professional autonomy. An association or a college of physicians, for instance, may issue practice guidelines. Peer review is a widely accepted way of monitoring physicians by physicians themselves. In these cases the individual autonomy is restricted but the professional autonomy is maintained. Because of this professional autonomy, physicians will probably prefer managed-care techniques designed or issued by the profession itself to techniques issued by a relative outsider, like a health insurer. Because of their individual autonomy, they will probably prefer control by incentives and control by persuasion or information to control by directives or authority. By means of directives or authority the physician's actions are restricted in an almost coercive way.

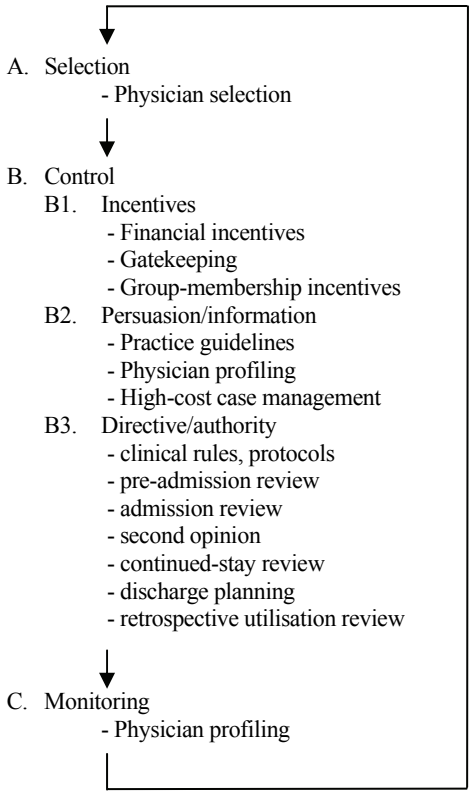
Finally, the monitoring phase of the agency cycle can also be identified in managed care. In the form of physician profiling two main goals are revealing information about the behaviour of the physician or about the outcome of the process to which the physician contributed, and informing the physician about his relative performance. Profiling can thus be viewed as a way to reduce the information gap between physician and third party by informing the third party as well as a way to reduce this gap by informing the physician.

Combining the agency or managed-care cycle, as pictured in figure 5.1, with the managed-care techniques described in this chapter results in figure 5.2.

Managed care requires a relationship between the third-party agent and the physician. Within this relationship the several techniques may be used in combination with each other. Although this is not the subject of this research, it may be hypothesised that the stronger are the relationships between both parties (towards integration), the less formal are the arrangements. The proximity of peers, group norms or the culture within a group of physicians may be very effective characteristics of such arrangements.

There is some evidence regarding the overall performance of organisations that apply managed-care techniques. This suggests that managed-care organisations provide health care at lower costs than fee-for-service indemnity plans. There is some ambiguity over the issues of quality of care and enrolee satisfaction though. Although most studies showed at least comparable quality, a few studies showed adverse results for some specific health problems. The lower satisfaction was mainly the result of the enrolee's restricted choice of physicians due to physician selection by the organisation.

Figure 5.2. The managed-care cycle unfolded



Less is known about the exact arrangements between third party and physician and about the effect of specific managed-care techniques in terms of cost, quality and accessibility of care. Often, the arrangements are complex and different per third party. Hence it is difficult to draw general conclusions and to infer the independent contribution of each technique to (the outcome of) health care, organisational performance et cetera. It is also difficult to infer whether the different techniques are substitutes or whether they are or can be used complementary. Finally, the focal point of many studies was the third party instead of the physician. Such studies are informative as it is the third party that may apply the techniques, but in the end it is the physician's behaviour that is of interest. As the physician may have contracts with several third parties, he may be confronted with different controlling en monitoring techniques. This combination of techniques will influence his behaviour. Therefore, studies that have the physician as focal point are at least as important.

As explained in chapter 4, outcome uncertainty is an important characteristic of the relationships in which a physician is involved. The outcome in terms of the patient's health

status is uncertain and will only partially result from the GP's actions. It will result from the natural course of the disease, the medical treatment, the behaviour of the patient, and other health influencing factors. As a result, the third-party agent will have problems assessing the appropriateness of the physician's actions. Hence contracts that will specify (all) possible outcomes and that, for instance, relate payments to factual outcomes are not common in health care. Also the majority of managed-care techniques are focussed on the behaviour of the physicians. This does not alter the fact that the behaviour of the physician has some effect on the outcome. If the behaviour has no effect on the outcome in terms of the patient's health status, then it will at least have some effect on the outcome in terms of volume or costs of care, patient satisfaction, availability of scarce care for other patients et cetera. Monitoring and analysing claims data may reveal important information then. Indeed, some third parties adjust their payments to physicians on the basis of utilisation, cost or quality measures, consumer surveys, physician productivity or other measures.

As the outcome in terms of health status will be the result of many factors besides the GP's actions, it is difficult to make the physician (partly) responsible for a negative outcome. The volume or the costs of the care provided, however, are closely related to physician's actions. Hence the third party may design an incentive system that is closely related to the volume or the costs of care. An example of such an incentive system is financial-risk sharing by which means the third party shares the responsibility for the costs of care with the physician. The rationale for financial-risk sharing and the way such arrangements can be structured is subject of the next chapter.

6 FINANCIAL-RISK SHARING IN THEORY

6.1 Introduction

In the previous chapter we described several managed-care techniques and grouped them according to the triptych of agency theory, selecting, controlling and monitoring. A specific controlling technique is the use of financial incentives. One form of a financial-incentive system is financial-risk sharing between third-party agent and GPs. How risk-sharing arrangements may be structured financially, and how these arrangements relate to the organisational side of the relationship between both parties, is subject of this chapter. An answer is sought to the following research questions:

What is the rationale for financial-risk sharing between third-party agents and general practitioners?

and

How can systems of financial-risk sharing be structured?

In section 6.2 the rationale is discussed for systems in which the risk is shared between third-party agent and GP. In section 6.3 is analysed how such systems can be structured.

6.2 The rationale for financial-risk sharing

6.2.1 Introduction

We return to the concept of 'sharing' in the term 'financial-risk sharing' in subsection 6.2.4, but the other two parts of the term are defined more precisely here. The first part of the term indicates that the potential results of the risks involved are purely monetary by nature. Other risks, like the risk that due to under- or over-provision of care the reputation of a third party or a physician may be damaged, are left out of consideration here. The second part, 'risk', can be defined and classified in several ways. One option is to view risk as a potential deviation from a norm, for instance, an expected value. Then, risk can be divided as follows (see De Wit's interpretation of Härterich, De Wit 1994, p. 4):

1. pure risk, i.e. only negative deviations (loss);
2. speculative risk, i.e. positive as well as negative deviations (profit or loss);
3. profit side, i.e. only positive deviations.

In the literature on provider reimbursement the term 'risk' is used differently. Some authors use it to denote pure risk (see, for instance, Kongstvedt 1993a) whereas others regard risk to be based on a chance for profit as well as loss (see, for instance, Miller 1996).

Unless otherwise stated, here the term 'risk' refers to the speculative risk. This is a rather obvious definition in view of the frequent use of capitation systems in which for a risk-bearing (primary care) physician the financial results may turn out well or badly.

6.2.2 Insurance risk

In absence of some form of third-party insurance, an individual will have to pay for each health-care good or service delivered by a physician. As it is largely impossible to predict accurately the moment and the frequency of consumption as well as the nature and the amount of care, the individual faces a pure risk (see previous subsection). Introduction of a third party enables the individual to reduce this risk.[1] By paying an insurance premium, the individual (from then on the insured) substitutes a certain financial loss for the uncertain financial loss associated with the occurrence of illnesses. The third party's insurance function thus involves a risk transfer from the insured to the third party, i.e. the health insurer. The risk the third party now bears is labelled the *insurance risk* (Kirkman-Liff and Van de Ven 1991). It is defined as the possible (positive or negative) variation around a normative cost level – a level as reflected in the insurance premium. The pure risk as faced by the individual is thus changed into a speculative risk for the third party. On a free insurance market and with perfect information the premium will resemble the normative value, i.e. the expected costs per insured (risk-rated premium).[2]

For a third-party insurer holds that the expected loss is the basis for the cost price and the premium and that the variability around the overall loss determines the probability of a ruin.[3] Especially in case the causes of the losses are not correlated, the variability around the mean loss will decrease if the number of insured increases. The variability around the overall loss will increase then, but in general the standard deviation of the overall loss is smaller than the sum of the standard deviations of the individual losses. Is the group of insured infinite, then the variability around the mean loss will even approach zero and the expected mean loss will resemble the mean loss on a national level (law of large numbers). Further, the losses will be distributed normally (central limit theorem) (Voûte 1987).

Once responsible for the insurance risk, the third party has to reimburse a patient for the use of health services (a reimbursement system) or pay the physician directly for the care provided (a contract or integrated model; see subsection 2.3.2). Ideally, the payments would be directly related to the costs the third party would consider being 'necessary' or 'needed' to treat the patient. In other words, the third party would only experi-

[1] If third-party insurance is absent, the physician runs the risk that his income drops considerably due to a lack of 'clientele', or that after the delivery of goods or services the patient turns out to be insolvent. Introduction of third-party insurance has the additional advantage that it takes away the risk of a patient being insolvent.

[2] By means of premium regulation and for reasons of solidarity, however, a regulator may force the third party to set a premium that more closely reflects mean costs per insured.

[3] The probability of a ruin is the probability of a loss in one year of a certain (agreed upon) amount of money, for instance with the size of the stock and/or the reserve capital, resulting in an inability to meet one's liabilities (Tolley et al. 1987; Seal 1969).

ence those costs that have to be made to provide an optimal amount of care, that is, a level of care given 'perfect agency'. This optimum would imply:

A. The optimal diagnosis: timely and correct diagnoses instead of missed, unnecessary or wrong diagnoses (for instance, no upcoding, i.e. choosing diagnoses which yield larger marginal benefits to the physician than the proper diagnoses would).
B. The optimal treatment: no unnecessary or unnecessarily expensive treatments and no postponing or withholding of treatments (may result in higher future costs).

Furthermore, the physician-as-perfect-agent would use preventive medicine and patient education to reduce initial physician visits in order to lower the costs associated with the insurance risk. The definition of the insurance risk, the possible deviation from a normative cost level, is thus made under the assumption of perfect agency. For several reasons, however, it is hard, if not impossible, to reach this optimal level of care. As a result, the third party runs a second risk: the risk of imperfect agency.

6.2.3 Risk of imperfect agency

Besides the insurance risk, which results from the stochastic nature of the occurrence of illnesses, the risk transfer from insured to third party involves a second risk. Kirkman-Liff and Van de Ven (1991) labelled it the *risk of the provision of cost-ineffective care*. It consists of duplication of tests and the provision of services that are not strictly necessary from a medical point of view, inappropriate care, care provided by overqualified providers, et cetera. In addition to the insurance risk and the risk of the provision of cost-ineffective care, the risk of underprovision of care is also of importance here. The third-party agent is not only interested in reducing cost-ineffective care, but also in reducing underprovision as this may harm the patient and may also result in higher future costs. As both are agency problems, the risk of the provision of cost-ineffective care and the risk of underprovision of care together form the *risk of imperfect agency*. Is the insurance risk the variation around a normative level of costs given perfect agency, the risk of imperfect agency is the variation around the insurance risk as a result of a physician acting as imperfect agent for the patient.

There are several reasons why the actual level of care may deviate from the optimal level of care as mentioned in the previous subsection. Some causes stem from the patient-physician relationship as such, whereas other causes result from the presence of third-party insurance. The first group of causes consists of agency-like problems within the patient-physician relationship (see chapter 4). Some of these problems are information problems, whereas other problems result from conflicting interests. The asymmetry of information between physician and third party allows the under- or over-provision of care. A physician may, for instance, be susceptible to the incentives emanating from the payment system. In fact, the risk of imperfect agency may for an important part be the result of the difficulty of establishing a perfect payment scheme. As it is virtually impossible to define for each state of health and for each patient the optimal level of prevention, the optimal diagnosis and the optimal treatment, and as it is therefore equally difficult to develop an accompanying payment system, the third party is forced to use a less

than perfect payment scheme. If, for instance, the physician's marginal benefits (MB) of providing additional care exceed his marginal costs (MC), then a fee-for-service system stimulates him to provide additional goods and services. If the physician's marginal costs of treating a patient are higher than his marginal benefits, which may be the case with capitation payments, then the payment system stimulates the physician to reduce his efforts in delivering care and to refer the patient to other, and maybe more expensive providers of care.

The second group with causes of imperfect agency contains two forms of insurance-induced moral hazard: consumer-induced moral hazard and supplier-induced moral hazard. May the agency problems within the patient-physician relationship result in under- as well as over-provision, the result of moral hazard will be over-provision of care.

Table 6.1. Causes of the risk of imperfect agency

A. Patient-physician relationship:
1. Information problems:
i. Inadequate experience or limited information physician
2. Conflicting interests:
i. Incentive for physician to provide additional, unnecessary, care (for instance, if MB > MC)
ii. Incentive for physician to underprovide care, or to refer patient on medically unnecessary grounds (for instance, if MC > MB)
B. Moral hazard:
1. Consumer-induced moral hazard
2. Supplier-induced moral hazard

The patient is primarily responsible for the occurrence of consumer-induced moral hazard. The third party may try to control this type of moral hazard by means of cost-sharing arrangements. However, consumer-induced moral hazard is also related to the physician's agency role and so may contribute to the risk of imperfect agency. Firstly, once the patient has entered the medical circuit, it is the physician's job to restrain him from the so-called moral hazard *ex post* – the demand for more or more expensive care. Secondly, it may be argued that the physician-as-agent should also try to influence initial physician visits by means of preventive medicine and patient education – the reduction of moral hazard.[4]

6.2.4 Dealing with both risks

Once a third party accepts an applicant, it becomes responsible for both the insurance risk and the risk of imperfect agency. In theory, the third party has a choice then between four strategies to handle these risks: risk bearing, risk shifting, risk splitting, and risk sharing.

[4] See Horgby (1995, p. 31) for the distinction between moral hazard *ex ante* and moral hazard *ex post*.

Risk bearing
A first strategy for a third party would be to bear the risks and to spread it over all policyholders by means of a higher premium (risk premium as well as loading fee). As long as an insured wants to pay for it, a third-party agent could draw up a policy that would allow the insured to consume and physicians to provide care freely and according to their own view (a 'Total Freedom policy', Van de Ven 1996). This, however, would be a charter for over-provision of care. The insured's freedom of choice would prevent the third party from contracting physicians selectively and would allow the insured to visit the physicians he prefers, even if they provide substandard quality or cost-ineffective care in the opinion of the third party. The insured's freedom of consumption would hinder the third party to influence the nature and the amount of care, where and by whom it is delivered, et cetera. This may result in lower quality of care due to, for instance, false test results and iatrogenesis. It is in the insured's interests, therefore, if the third party also offers a policy with conditions that allow the third party to manage the care (an 'Appropriate Care policy', Van de Ven 1996).[5] Notice that the risk-bearing strategy is inconsistent with the concept of the third-party agent.

Risk shifting
A second strategy for a third party would be to shift the risks to independent physicians or to a middle tier, like a provider organisation. On the one hand, this would remove the charter for over-provision of care. On the other hand, the physicians would bear full responsibility then, which implies responsibility for the insurance risk. It is the insurer, however, who should bear this risk; it is one of the rationales for the presence of a third party. The insurer is capable of dealing with this risk as the total number of insured is large – making it possible to use actuarial techniques based on the law of large numbers – and because he has the means at his disposal to meet fluctuations in costs and to compile

[5] It should be noted that an 'Appropriate Care policy' is not inconsistent with the concept of agency. At the moment an individual chooses an 'Appropriate Care policy' instead of a 'Total Freedom policy', he will probably do so because he wants the third-party agent to reduce the amount of cost-ineffective care delivered to himself and to other insured. Once the insured actually needs health care, his preferences may change (see also the 'veil of ignorance' in subsection 4.4.3). Although he may still be unwilling to pay a premium for cost-ineffective care provided to other insured, he now may want to consume such care himself being convinced that it *may* be effective. A third party is not an imperfect agent if it sticks to the insurance contract both parties once concluded and refuses to reimburse this care.
Some argue that managed care in general and the limitation of the insured's freedom of choice in particular may threaten the quality of care, which seems to be in contradiction with the above. Ohsfeldt et al. (1998) gave some arguments used by proponents of 'freedom of choice' laws. Firstly, freedom of choice allows patients to choose other physicians if they suspect the preferred physicians of quality skimping. Secondly, quality may be enhanced by a reduction of the time a patient has to travel. Thirdly, such laws may prevent that solely large chains or groups of physicians are contracted. Here, however, two assumptions are made. One assumption is that an individual can choose between a 'Total Freedom policy' and an 'Appropriate Care policy'. If he regards a limited freedom of choice as undesirable, then he can opt for the 'Total Freedom policy'. The second assumption is that a third-party agent is stimulated – for instance, by a regulator – to avoid quality-skimping physicians and to guarantee a sufficient spread of physicians, and that small providers are contracted if that is in the interests of the insured.

and analyse large sets of claims data. Moreover, bearing the full risk may prompt physicians to take, from a third-party agent's point of view, undesirable measures, like cream skimming or quality skimping.

Risk splitting

Another option for a third party to deal with the risks would be to choose a middle coarse between bearing all the risk and shifting all the risk. As noted, it is a function of the third party to deal with the insurance risk. Also argued is that the risk of imperfect agency is to a large extent under control of the physicians (see also table 6.1). An obvious solution, therefore, would be to split the risk: bear the insurance risk but shift the risk of imperfect agency to the physicians. Bearing this latter risk would stimulate an individual physician to gain more experience and to increase his knowledge if this could reduce the risk. Moreover, if many physicians were risk-bearing, an extra effect would be that the medical profession as a whole is encouraged to increase collective knowledge of diseases and medical practices. Further, bearing this risk would decrease the incentive to provide unnecessary care or to demand for unnecessary care to be delivered by another provider of care. The problem, however, is that it is difficult to separate the insurance risk from the risk of imperfect agency. As noted in the foregoing, it is virtually impossible to define, in advance, all possible states of health and disease with the accompanying optimal diagnostic and treatment patterns. Even retrospectively it will be hard to decide whether the physician's actions were optimal. It will be difficult for a third party to measure the outcome of the physician's practising. And even if measurable, it will be difficult to decide whether the outcome results from the physician's efforts, from other providers' contributions, from the patient's behaviour, from other health-influencing factors or from the natural course of the disease. Hence an unsatisfactory outcome can not easily be ascribed to the physician's behaviour. Further, a satisfactory outcome may indeed be the result of the physician's actions, but it will be hard to state whether the same outcome could not have been obtained in a more cost-effective manner. These problems imply that the middle coarse of bearing the insurance risk and shifting the risk of imperfect agency to GPs is practically not feasible.

Risk sharing

A more feasible middle way for a third party to deal with the insurance risk and the risk of imperfect agency seems to share the risks with physicians. Although a second-best option – the physician becomes partly liable for the insurance risk – it is a seemingly satisfying strategy between bearing all the risk and shifting all the risk. In a financial-risk sharing arrangement surpluses and deficits in the budgets for certain types of care are distributed among third party and physician. The physician's responsibility is thus larger than zero but less than hundred percent. Such an arrangement offers incentives to reduce the provision of cost-ineffective care, which can be balanced with the protection of a physician from too much risk as well as the attempt to prevent him from taking undesirable measures. Furthermore, the physician will not only be stimulated to reduce overprovision. The incentive for underprovision will be less in comparison with the risk-shifting option. Moreover, by means of a well-structured arrangement it will be made

unattractive to postpone care, to shift the liability to other providers, to reduce the quality of care, et cetera.

In a risk-sharing arrangement between a third party and a physician, the risk that was transferred from the insured to the third party is now partly transferred to the physician. Shifting a part of the financial risk from first party to second party via a third party may seem a circuitous route, for the individual could also have passed the risk directly to the physician by paying him a monthly or yearly premium. In fact, the individual would then become subscriber to physician services and the physician would become insurer. The insurance principle of spreading risks based on the law of large numbers, however, would demand large practice populations and might force the physician to use risk-management techniques, like risk analysis, risk reduction, and transfer of risks. This would especially be true if the physician-as-insurer would bear full responsibility for a large package of goods and services. The circuitous route, therefore, is an effective way to spread the risks more evenly and to protect the physician from too much risk, and to save him the bother of product development, marketing, management et cetera.

The third party can influence the location of the financial risk by means of the payment system. Jegers et al. (2002) distinguished between fixed versus variable payment systems and between retrospective versus prospective payment systems (see also subsection 5.4.1.1). In case of a variable, retrospective payment system, the third party will bear the risk. In case of a fixed, prospective payment system the physician will bear the risk.

6.3 Potential effects of financial-risk sharing

Financial-risk sharing is one of the third party's potential answers to the agency problems of differential information and conflicting interests within the patient-physician relationship and within the relationship between third party and physician. It aims at shifting a part of the financial risk to physicians. As argued, the ultimate goal of risk sharing is to stimulate physicians to provide high-quality and cost-effective care. Risk-sharing arrangements are a compromise between risk bearing and risk shifting. In case the third party shifts the risk, the physician faces strong incentives for efficiency but also for undesirable behaviour. In case of risk bearing, the incentives for undesirable behaviour are low but so are the incentives for efficiency. There is thus a trade-off between efficiency and high-quality care.[6]

High-quality and cost-effective care

Well-designed risk-sharing arrangements stimulate GPs to reduce the costs of care, while maintaining or even improving quality. The physician is, in other words, stimulated to reconsider several aspects of the delivery of health care. Firstly, he may reconsider *which*

[6] Newhouse (1996) introduced the trade-off between efficiency and selection in the context of reimbursement of health plans. Besides selection, a physician can show other kinds of undesirable behaviour too.

care is to be delivered. A wait-and-see policy may be satisfactory. If not, he has to decide whether treatment should be palliative or curative. Extensive curative care may not be a cost-effective option if the patient is incurably ill. Further, the question arises whether the patient's complaints are somatic, psychosomatic, or mental. A second point of reconsideration may be *who* should provide the care. A medical specialist, a GP, a paramedic? Can a practice nurse substitute for a GP? Thirdly, *where* should the care be delivered? Care may be provided in a hospital, on an inpatient or an outpatient basis. But maybe quality can be maintained or even enhanced if care is provided within a relatively lower-cost setting, like the community. A fourth question is with *what intensity* (how often and how long) care should be delivered. Additional diagnostics may be unnecessary, and the marginal value of, for instance, twelve treatments to nine may be low.

The effectiveness of risk-sharing arrangements, however, will stand or fall on their financial and organisational structures. With too little risk, the physician may not respond to the incentives and the third party's aims of high quality and cost-effectiveness may not be achieved. Too much risk, on the other hand, may provoke undesirable behaviour, like cream skimming, cost shifting, and quality skimping.

Cream skimming

Besides the use of strategies to enhance the quality of care and to improve the cost-effectiveness, other, but undesirable behaviour may also be induced. Once a GP is financially responsible for a part of the future costs for his practice population, he may be tempted to reduce his expected costs by means of cream skimming. The term cream skimming originates from the insurance industry and refers to the selection by an insurer of low-risk (or: preferred) insured, i.e. insured for whom the expected costs are lower than the matching premium. It may occur if the insurer has information about the insured's expected costs that is not fully reflected in the premium. If an individual physician is unable to adjust the height of the reimbursement to the (expected) costs of care for a certain patient, he may face an incentive for cream skimming too. This may be the case with prospective payment systems (like capitation) but even with 'cost-based' reimbursement (like fee for service) if the marginal costs exceed the marginal revenues. Applied to the practice of a GP, cream skimming is defined here as the selection by a physician of low-risk (or: preferred) patients, i.e. patients for whom the expected costs (as predicted by the physician) are lower than the reimbursement.[7] He may attract low-risk patients or, given a certain practice population, try to avoid (dump) high-risk patients.

When it is an attractive option to select patients according to their expected costs, a physician has several techniques at his disposal. Some of these techniques resemble those that can be used by an insurer, other are specific for the practice of a GP. The location of his practice (in a region with a relatively large proportion of low-risk patients) as well as

[7] Some define creaming by providers in a different way, namely as the over-provision of services to low cost patients (see Ellis 1998, p. 538). This, however, is a somewhat restricted definition for the provision of services is not a necessary condition for skimming. In case of capitation payments, providers face at least two incentives. The first is indeed to attract patients with expected lower costs than the capitation payments. The second, however, is not to *over-provide* but to *underprovide*.

its accessibility (little public transport, a building with stairs) may help the physician to attract low-risk patients. Further, he can advertise selectively the practice's opening hours, not specialise in chronic or other expensive diseases, provide subtly poor-quality care to high-risk patients (waiting times for making an appointment as well as in the office) et cetera (Kirkman-Liff and Van de Ven 1991). Other strategies are to avoid contracts with insurers having a relatively large amount of high-risk patients, or to inform the patient with a high-cost disease of other providers 'who are more capable of dealing with such problems'. If a GP has to purchase follow-up care on behalf of his practice population (like in a GP fundholder system) he may contract providers selectively. No or only little contracts are concluded then with providers specialised in high-cost care.

Cost shifting

A major characteristic of the GP is his referral function: he may send a patient to another provider of care who is more qualified or authorised to deal with the patient's complaints. It depends on the structure of the risk-sharing arrangement whether this *care shifting* also results in *cost shifting*. If a physician is financially responsible for specialist care, referring a patient to a medical specialist only implies the shifting of care. Cost shifting occurs if the care delivered by the provider to whom the patient is referred is not part of the GP's risk package (the package of care for which the GP bears financial responsibility, see subsection 6.4.2).

If the GP is liable for only a part of the patients' care, and if it is not exactly clear to others (like the third party, his colleagues or the patient) which care or which provider is indicated (i.e. if there is an information asymmetry), he may use cost shifting as another strategy to reduce his costs. Cost shifting is considered to be improper if, from a medical and economic point of view, the patient should have been treated by the GP himself or by another provider for whose care the GP is financially responsible but, nevertheless, is referred to a provider outside the risk package. An informational advantage from the physician over the other parties is, however, not an essential condition for cost shifting, for it may work equally well within the so-called 'zone of uncertainty' (see also subsection 4.2.2). For example, it may be unclear whether, given a certain diagnosis, a patient should be treated within a hospital or not. If the physician is responsible for community care then but not for hospital care, cost shifting can be an ideal cost-reducing strategy.

Quality skimping

For patients not avoided by means of cream skimming and not referred to another provider, a GP may try to reduce his costs by skimping on the quality of care. Main possibilities are to reduce his efforts while treating a patient, to postpone necessary care, or even to withhold it. Skimping on quality may reduce costs at short notice but it can be a highly uneconomical strategy in the long run. Characteristic for primary care is the long-term relationship between patient and physician. Therefore, postponing or withholding care may deteriorate the patient's health and result in higher costs later on. Furthermore, postponing and withholding care for financial reasons is not likely to occur if the physician is paid on a cost-basis. It may, however, be a problem in case of prospective payment systems – of which the use is more likely in case of long-term relationships – espe-

cially if the payments are not adjusted for health. For that reason, quality skimping seems to be a more attractive option for risk-bearing medical specialists, as their relationships with patients is generally of relatively short duration (and payments are often on a fee-for-service basis). Nevertheless, a GP may use skimping on quality too, particularly if he does not expect costs to be higher in the future or as a strategy to discourage high-risk patients.

Except for the length of the relationship, the physician is constrained in his ability to skimp on quality by, for instance, the extent of the information asymmetry, his reputation and the extent of competition. The better patient and third party can monitor the physician, the more difficult skimping will be. Fear for his reputation may deter the physician from quality skimping, although this is related to the asymmetry of information and the risk of detection. If patients are able to 'vote by feet', the physician may avoid skimping. On the other hand, skimping may be used as a tool for cream skimming and the voting may be encouraged then.

6.4 The structure of risk-sharing arrangements

6.4.1 Introduction

Once is decided upon risk sharing as incentive system to control the GP, a contract has to be devised arranging the financial as well as the accompanying organisational relationship between both parties. Such contracts can be structured as follows. A first matter to be resolved is for which care the GP bears (some) financial responsibility (see subsection 6.4.2). A second matter is the determination of the proportion of the physician's practice population that is covered by a risk contract (subsection 6.4.3). Next, both parties have to agree upon a norm with which the physician's behaviour, or the outcome of this behaviour, can be compared (subsection 6.4.4). Depending on the terms of the contract, a deviation or a meeting of the norm will have a financial consequence for the physician. The physician may, for instance, be given a bonus (subsection 6.4.5). To limit the impact of the incentive system on the GP (i.e. to protect him from to much risk), additional measures may be added. These measures may affect the maximum of the bonus as well as the malus (subsection 6.4.6).

The aforementioned aspects of a contract can be viewed as 'variables'. By altering them, the incentives the GP faces vary accordingly. Potentially, this could result in a change of the physician's behaviour and the outcome of his behaviour.

6.4.2 Risk package

Crucial in structuring a risk-sharing arrangement is the composition of the risk package, i.e. the scope of goods and services for which costs the GP bears at least some financial responsibility. It is noted before that due to his role as gatekeeper and co-ordinator of medical services, the GP has a considerable influence on the nature, quantity and quality of goods and services to be delivered by other providers of care. This influence justifies

the extension of his financial responsibility by including care beyond the primary care services in the risk package.[8] The risk package may be extended then to include certain referral care (specialist services, like consultations, surgical procedures, anaesthesia and obstetrics), hospital care (outpatient as well as specific inpatient care), and ancillary services (like laboratory, radiology, physical therapy, and pharmacy).

In designing the risk package, third party and physician have to take the following into consideration. Firstly, the probability that in a certain period the physician incurs costs for one type of care will be higher than for another type. Moreover, given costs in that year, the variability in costs for the different types of care may differ as well. The type of care included in the risk package thus determines the insurance risk the physician runs. The larger the insurance risk, the more (costly) risk-reducing measures are required to protect him.

A related question is whether it is the third party or the GP who has to purchase care. In a system in which the physician, besides being financially responsible, also has to draw up and conclude contracts with other providers of care – this is comparable with GP fundholding in the UK – it is important that he is able to judge the referral care economically as well as medically. For that reason, in the fundholding system only standard, relatively inexpensive care without 'open-ended treatments' for which the physician would be able to diagnose the case and to estimate its costs was included (Glennerster and Matsaganis 1993).

A third matter to be taken into consideration is that the composition of the risk package may affect the occurrence of cream skimming and cost shifting. If the physician bears no financial responsibility at all (or only for his own primary care) there is no (or only a small) incentive for cream skimming, for expenses are reimbursed or patients can easily be referred to other providers of care. If the physician is responsible for the costs of all care, then cost shifting is no longer possible and the incentive for cream skimming is maximised. Between these extremes it will depend upon the care excluded from the risk package whether cost shifting is an option. To shift the costs to another provider, the care delivered by this provider should not only have been excluded from the risk package but also substitute for the care included in the physician's risk package.

Fourthly, once the risk package is established, it has to be decided whether this package is considered to form a whole or whether it is divided into separate cost categories. The advantage of categorising health-care costs is a better insight into the cost structure, making it easier to track changes and to identify room for improvement. Common is a functional categorisation in inpatient services, outpatient services, physician services, other medical services, ancillary services, prescription drugs, reinsurance premiums, and other medical costs (Blox et al. 1989, Ward 1993). Another main advantage of a division into separate cost categories is that it permits the third party to vary the extent of the GP's responsibility, i.e. the extent of financial-risk sharing, per cost category (see also subsection 6.4.5). Disadvantage of the division into cost categories is the inherent danger of reduced substitution, or even undesirable substitution of expensive care for less expen-

[8] Whether the inclusion of certain care in the risk package is considered an extension, depends on one's definition of primary care services. Definitions may differ, for instance, per country.

sive care if the physician bears more risk for the relatively inexpensive goods and services. The division may also result in increased administration costs.

A final matter mentioned here is that the composition of the risk package may also influence the behaviour of other providers of care. Firstly, these providers may have been made jointly responsible by participation in a risk pool.[9] Secondly, if not participating in the risk pool, they may react to the altered behaviour of the GP. For instance, if GPs under a risk contract substitute primary care for secondary care, secondary care providers may try to compensate their loss of income by increasing the number of services that are not parts of the risk package (Delnoy and Stokx 1993).

6.4.3 Size of the practice population (at risk)

The foregoing may apply to a one-patient contract, but it is evident that in practice a contract will apply to a larger part of a physician's practice population. Which proportion of his practice population is covered by a risk contract is of importance for at least three reasons. Firstly, the number of patients for whose care the physician is financially responsible determines the magnitude of the incentives and presumably the effect of the risk contract. All other things being equal, the larger the proportion, the more effect the incentives emanating from the contract will have. Obviously, an incentive system will have a stronger influence on the physician's behaviour if it applies to, for instance, 800 of his 2500 patients than if it applies to 80 of this 2500.

There is a second reason why the proportion of a physician's practice population covered by a risk contract is important. The larger this proportion, the more difficult it is to shift the costs incurred under that contract to other third parties with which the physician has no risk contract. An example of how this might be done is to compensate a loss of income due to the risk contract by increasing the number of claims for patients for whose care he is not financially responsible.

A third reason is that the larger the group of patients for whom the physician has concluded a risk contract, the less vulnerable he is to random fluctuations in the health-care costs of his patients (i.e. the insurance risk).

Which proportion of his practice population is covered by a risk contract will depend on the number of third parties the physician is confronted with.

One third party

A GP may have a relationship with a single third party. This will be the case in a health-care system with just one third party (like a National Health Service), if a third party has a regional monopoly, or if the physician has an exclusive relationship with the third party (like within a closed-panel or staff HMO). As the physician sees only patients who are insured with the third party in question, it is likely that, in principle, all of his patients are covered by the risk contract. Some of them may be excluded though. Decisions about the exclusion of members may, for instance, be based on their health status. Some patients,

[9] A risk pool is a group of providers who share in the rewards and penalties from surpluses and deficits in the budgets for certain types of care (see subsection 6.4.6.4).

like those with AIDS or those who are in need of a transplant, may be excluded because of their expected high costs.

Several third parties
Alternatively, the GP may practise within a health-care system with several third parties. The market share of each third party as well as with which third parties the physician has a risk contract determines the ultimate proportion of his practice population that is covered by such a contract. Again all members of a third party with whom the physician concluded a risk contract may be covered, or some of them may be excluded because of their health status.

The presence of several different risk contracts may result in some unforeseen behaviour. Theoretically, the incentives resulting from the several risk contracts may balance out. If one-third part applies a cost-increasing bonus and the other a cost-decreasing one, the effect (supposing comparable proportions of members) may be a middle course. Another effect may be that if one contract is dominant, the physician will act on it. Attempting to change his behaviour for a minor proportion of his patients may not be worthwhile then.

6.4.4 Normative level of care

Definition and function of a norm
At the heart of a risk-sharing arrangement lies a norm. The third party can establish a normative level of care with which the actual level is compared. The outcome of this comparison can form the basis for ancillary payments in addition to the basic payment methods. A norm can be defined as 'an official standard or level of achievement that you are expected to reach or conform to'.[10] In this thesis, a norm is thus a standard or a set level of achievement for the GP, although it will depend on the terms of the contract whether the norm functions as a target or as a threshold. In case of a target, the norm is indeed a level of achievement that the physician is expected to reach or conform to. In case of a threshold, the physician should attempt to keep below or – again depending on the contract terms – above the norm.

A norm may be specified as a certain quality level, like compliance with guidelines or protocols. Other options are to specify a norm as a certain volume of care, like a number of prescriptions or referrals per patient per year, or as a cost level. In the first case, the physician can only influence the financial risk by influencing the amount of care. In the latter case he may try to influence the volume as well as the price of care, or attempt to substitute relatively inexpensive care for relatively expensive care. As the *financial* risk is the starting point in a risk-sharing arrangement, the normative level of care will probably be expressed as a cost level.

Setting a normative level is a critical part of designing risk-sharing arrangements. The norm determines the GPs' compensation as well as the incentives they face. Such arrangements, therefore, should meet at least two requirements. First of all, the normative

[10] Collins Cobuild English Dictionary 1995, p. 1122.

level for physicians should be such that each physician has in principle the same possibility to meet the norm. Even if physicians would act as perfect agents and provide the optimal diagnoses and treatments, factors exogenous to the physicians may make it virtually impossible to conform to the norm. Important, for instance, are the patient characteristics. As these may differ per practice, the norm should account for the resulting cost variations. In this way a settlement takes place of systematic differences in health status of the practice populations. This is the requirement of *fairness*. Secondly, the norm should provide the physician with the proper incentives. Within the proposed framework, a third-party agent applies a financial-incentive system with a norm to stimulate the physician to act as agent for his patients by providing cost-effective high-quality care. Obviously, such a system should not provoke adverse physician behaviour, like cream skimming, cost shifting, and quality skimping. This is the requirement of *incentive compatibility* (see also subsection 3.3.3).

The basis of a norm

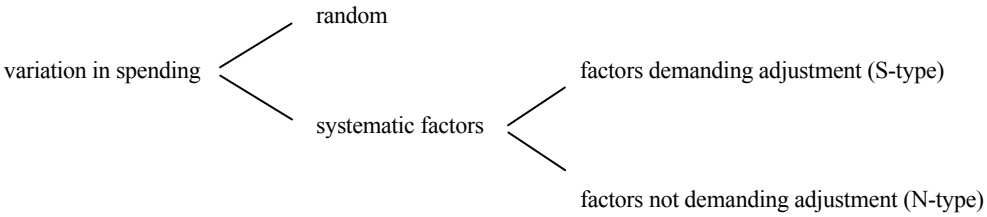
The main question is upon what the norm should be based. As explained in subsection 6.2.2, the ideal normative level would equal the costs that would be considered as 'necessary' or 'needed' to treat a patient given a right diagnosis. Van de Ven and Ellis (2000, p. 767) called such costs 'acceptable costs'. These costs were conceptualised as '(…) those generated in delivering a "specified basic benefit package" containing only medically necessary and cost-effective care' and were considered to be needs-based. They used the term in a different – though considerably comparable – context, namely in a discussion about the calculation of premium subsidies for health plans.[11] As well as for health plans, however, it is virtually impossible to determine such an optimal level of care and the related necessary costs for physicians. As a result, the third party will have to use other figures as yardstick.

Instead of using necessary or acceptable costs, an obvious way to determine a norm is the use of actual (or observed) costs. The problem with actual costs is that these costs differ per GP. After correction for practice size, substantial variations can be found. The variation in per capita spending per physician can be explained in two different ways. Van de Ven and Ellis (2000) explained variations in observed health care expenses across insured by dividing the variation into a random component and a systematic component. The random component is the result of the largely stochastic nature of health problems, which makes costs of an individual person largely unpredictable. The remaining variation results from several systematic factors. They mentioned seven classes of risk factors explaining the variation in spending across individuals. These can be grouped into characteristics of insured, of providers, of the region, and of the health plan. Three factors are characteristics of insured, namely age and sex, health status, and socio-economic status. Characteristics of providers, like practice style and supply, form another factor. A characteristic of the region where the care is provided are the input prices. A sixth factor is the

[11] A health plan is a kind of third party and is defined as 'a risk-bearing entity that performs at least some insurance function – i.e. it bears some or all of the financial risk associated with the random variation in health expenditures across individuals.' (Van de Ven and Ellis 2000, p. 758).

market power of the health plan, i.e. its ability to negotiate lower prices. Finally, features of the benefit plan may determine health care costs. Examples of such features are cost sharing with insured, contents of the benefit package, use of managed-care techniques et cetera.

Figure 6.1. Explanation of the variation in health care spending



Van de Ven and Ellis argued that, for reasons of solidarity, society may desire adjustment of the premium subsidies for health plans.[12] Only for some of the systematic factors, the so-called S-type risk factors, solidarity may be demanded. These will be the factors of which the resulting costs can not be influenced by the parties involved (i.e. third party or insured) and are 'acceptable' to be subsidised. Age, sex, and health status are typical S-type factors. Unless decided otherwise, the effects of the remaining risk factors, the N-type factors, are for account of the third party and may be reflected in the premium contribution to be paid by the insured. If, for instance, third parties have the legal and technical possibilities to influence the providers' practice styles, then adjustment for this factor may be considered undesirable. It depends whether a particular factor is designated as a S-type or as a N-type factor. One society may have other preferences for this than another. Further, it depends on the party in question. A factor may be considered N-type in the calculation of a subsidy for a third party, but S-type while determining a physician's norm.

The above mentioned explanation of Van de Ven and Ellis is related to the division of costs into the costs resulting from the insurance risk and those resulting from imperfect agency. Ideally, the variation amongst physicians in the average cost per patient would only reflect the positive or negative variation around the normative cost level, that is, would only reflect the insurance risk. The costs resulting from the insurance risk can be subdivided into a random part and a systematic part. A large part of the variation is ex ante random. The remaining part is determined by the S-type systematic factors. These risk factors cause variable costs that can be considered 'necessary', 'needed' or 'acceptable'. But as argued, information problems and conflicting interests may result in certain provider behaviour (N-type systematic factors) that may lead to additional costs: the costs of imperfect agency.

[12] A premium subsidy for a health plan is not necessarily the same as the norm for that plan. In some health insurance systems – the Dutch system is an example of this – the premium subsidy equals the norm minus a fixed amount. To cover this deficit, the third party may charge the insured a premium contribution.

The use of cost figures

To determine a normative level, health care costs can be used in several ways. The simplest way is to base the norm on the physician's historical costs. Overall costs per physician or figures on diagnostics, treatments, prescriptions and referrals per physician are needed then. The use of historical costs has an important advantage: the transition from a risk-free contract to a risk-sharing contract can be smooth because major disruptions of the physician's financing are absent. There are several disadvantages of using historical figures, however. First of all, it implies a continuation of an existing situation with possibly large differences in practice patterns. This is not a real problem as long as these differences are the result of differences in health status of the practice populations. It may be problematic, though, if the variations result from differences in provider characteristics. A norm based on historical figures means that the norm is adjusted for practice style, which may be considered undesirable. GPs who practised cost-effectively are punished, whereas their colleagues with a cost-ineffective style of working are rewarded with a relatively higher norm. A second disadvantage is that physicians face an incentive to increase the costs in the year before they conclude a risk-sharing contract. If historical figures are persistently used, then the incentives to increase costs, at least up to the normative level, remain. An additional problem may be the availability of detailed figures per physician and for the types of care to be included in the risk package. The conclusion is that a norm based on historical costs has the advantage of a smooth transition. However, it has the disadvantage that it accounts for N-type factors whereas it may not account for the S-type factors accurately. It will not lead to a fair compensation of physicians. Furthermore, it does not stimulate physicians enough to provide cost-effective care of good quality.

A more advanced way to determine a normative level is to use average per capita costs (averaged, for instance, per insurer, per region or per country). The advantage of using mean costs is that it is an easy way to calculate norms. It has several disadvantages, though. In the first place, the same objection as raised to the use of historical costs applies to this method: it implies the awarding of a potentially sub-optimal situation. Inefficiencies find expression in the average costs. Further, it is not a very accurate and fair way. A likely result is that some physicians are confronted with a norm that is actually too high and, therefore, have less opportunities to meet the required level of achievement and thus to get additional payments. Others may be favoured wrongly. Moreover, a norm based on average costs will inevitably result in the existence of groups of high-risk patients: patients for whom the physician expects the actual costs to be higher than the norm predicts. As a result, cream skimming becomes an option and may be an appealing strategy. Generally, this danger increases if more care is included in the risk package.[13]

Calculation of the norm can be improved by adjusting for one or more risk factors, like age and sex. The question is for which factors the norm should be adjusted. The more the model is refined, the more likely it is that its predictions will tally with the actual costs, and the more equal the possibilities are for physicians to become eligible for additional payments. Furthermore, a more refined model will make it more costly to

[13] This depends on the type of care included in the risk package (see subsection 6.4.2).

practise cream skimming. Providing cost-effective quality care becomes a more obvious option then. On the other hand, each further refinement will make the model more complex and more costly.

Risk factors
The systematic factors that can be used to calculate norms for GPs may differ from those used in the calculation of subsidies for third parties. As to the latter, age, sex, and health status were mentioned above to be typical factors for which adjustment may be demanded. These factors may also be used for physician norms because of differences in the distribution of such patient characteristics per practice, although with the marginal note that use of the factor health status is arguable. Strictly speaking, if the view is taken that only factors may be used which GPs can not influence, then its use is questionable. Except for new patients in their practices, physicians can exercise influence on their patients' health by means of patient education and preventive medicine, diagnostics, treatments et cetera. However, because of the requirements of fairness – the physician's influence on a patient's health is limited – and incentive compatibility – no adjustment for distinct differences in health status may provoke adverse physician behaviour – the factor health status is inevitable. The exact weight of the factor may, nevertheless, be up for discussion.

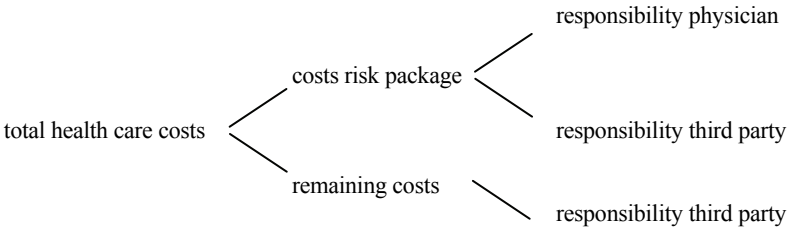
Besides direct patient characteristics, other systematic factors explaining the cost variations may be used in calculating risk-adjusted norms. Variations in patients' costs that are not the result of direct patient characteristics ought *not* to be reflected in the norms *as far as* GPs can exert influence on them. It seems not reasonable to make physicians responsible for cost variations that they can not influence, like those caused by other providers and that are the result of negotiations between third parties and the providers in question. If indeed other systematic factors are used, then the earlier mentioned requirement of fairness implies a settlement of systematic differences in health status of the practice populations *and* of differences in the other factors in so far as the GPs can not influence them. Characteristics of the health plan (i.e. the insurer) can be left out of consideration. Characteristics of the region will probably be exogenous to the physicians, although large physician groups (purchasing groups) may have some influence on them. Most obvious are the characteristics of providers. In a system of financial-risk sharing, the results of inefficient behaviour of GPs themselves and of providers with whom they may have concluded contracts concerning volume and price of care are, in principle, their own responsibility. Such financial liability for their own policy of provider contracting is found in, for instance, a fundholding system.

6.4.5 Bonus, malus and withhold

Characteristic for the risk-sharing arrangements as discussed in this chapter is that the GP has a limited financial responsibility for a limited package of health care. The third party remains financially responsible for the rest of the costs of the risk package and for the care excluded from the risk package (figure 6.2). After composition of a risk package and after calculation of an accompanying norm, the normative costs can be compared with

the actual costs. Eventually, the physician's financial responsibility can find expression in a bonus. Like defined in subsection 5.4.1.1, a *bonus* is an amount of money paid to the physician in supplement to the basic payment method only if he has met certain requirements. Such use of just a bonus system as an extra payment limits the physician's risk to the profit side. The third party provides the physician with only positive incentives then. The bonus may be paid out of the savings, for instance realised in the budgets for follow-up care. Does the third party provide just negative incentives or does it provide negative incentives as well, then the physician faces a pure or speculative risk respectively.[14] Just like a positively appraised outcome may result in a bonus, a negatively appraised outcome may result in a *malus*: an amount of money to be paid by the physician for not meeting certain requirements or for exceeding a norm.

Figure 6.2. Division of responsibilities among third party and GP



Bonus and malus can be calculated by comparing actual and normative costs at the end of a budget period.

A first option then is to pay a percentage of the difference between actual and normative costs, in case the difference between both costs is regarded as positive. If the difference between both costs is regarded as negative, then a malus may be collected. We label this first option a proportional bonus or a proportional malus (the larger the difference, the larger the bonus or malus).

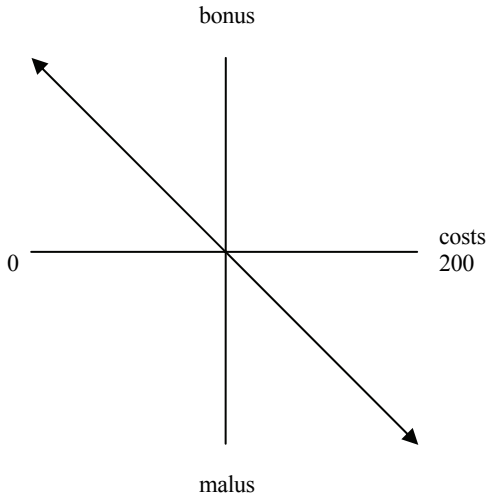
A second option is a bonus that is inversely proportional to the difference: the more equal actual and normative costs are, the larger is the bonus (or the malus). We label this second option an inversely proportional bonus or malus.

Another option is to use a fixed bonus or a fixed malus, irrespective of the difference between actual and normative costs. The difference between a bonus that is (inversely) proportional to the difference between both costs and a fixed bonus is that in the first the physician's behaviour is reflected more closely: the better he performs, the higher the bonus.

As mentioned in the previous subsection, the norm may function as a target or as a threshold. Its functioning is merely determined by the bonus and malus systems. This is illustrated by the following figures.

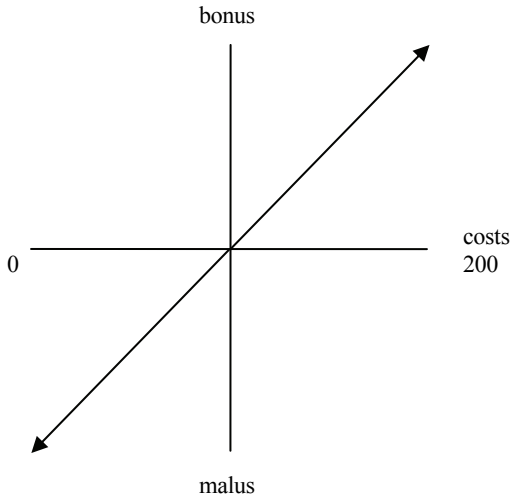
[14] As it is little motivating for physicians, the use of only negative incentives seems no option. Nevertheless, such systems are not uncommon on a macro level. By determining a macro budget for particular care, physicians are made responsible for cost overruns.

Figure 6.3. A cost-decreasing proportional bonus/malus system with threshold



In the bonus/malus system displayed in figure 6.3 the norm, which is 100, functions as a threshold. It provides the physician with an incentive to keep below the threshold by reducing health care costs. It is an example of a proportional system; the arrow indicates that the bonus (or malus) increases with the deviation from the norm.

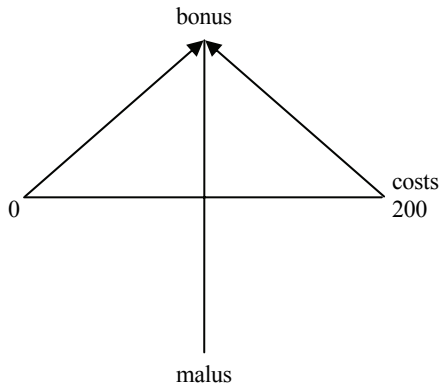
Figure 6.4. A cost-increasing proportional bonus/malus system with threshold



The system in figure 6.4 is also an example of a proportional system with a norm as threshold, but here the physician is stimulated to keep above the threshold. Fundamen-

tally different is the inversely proportional system displayed in figure 6.5 in which the norm functions as target. Here, the arrow indicates that the bonus increases if the deviation from the norm decreases.

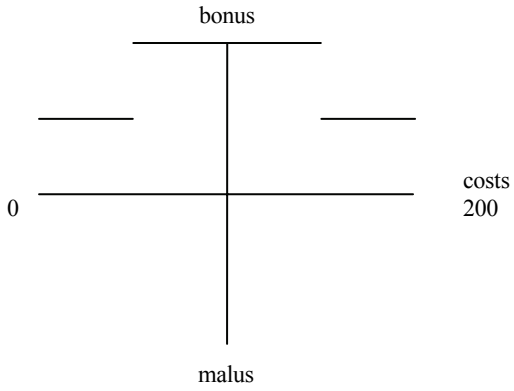
Figure 6.5. An inversely proportional bonus system with target



In principle, all systems may be usable and it will depend upon the goals of the third party which system is preferred. A third-party payer – under the assumption of relationships with physicians – may prefer the cost-decreasing proportional system with threshold. The inversely proportional system with target may be the third-party agent's choice. The reason for this is that in this system the desired level of achievement is made explicit: the normative level of care. The bonus is at its maximum if the norm is reached. As a result, the inversely proportional system with target reduces the danger of adverse physician behaviour. Whether this adverse behaviour consists of under- or over-provision depends on the bonus system. The cost-decreasing bonus system may result in underprovision, whereas the opposite holds for the cost-increasing bonus system. An interesting option then is to combine the incentive of the ancillary payment system with the contrasting incentive of the basic payment system. A cost-decreasing bonus system may be combined with a fee-for-service system, while a cost-increasing bonus system may be combined with a capitation system in order to balance the incentives.

In the cost-decreasing system the desired level of achievement is less explicit: a level somewhere below the normative threshold. Although it is probably not what the bonus system is intended for, such a system stimulates the physician to withhold care completely because the bonus is at its maximum then. Nevertheless, even a third-party agent may use the proportional systems displayed in the figures 6.4 and 6.5, for instance temporarily in order to decrease or increase the provision of a particular type of care.

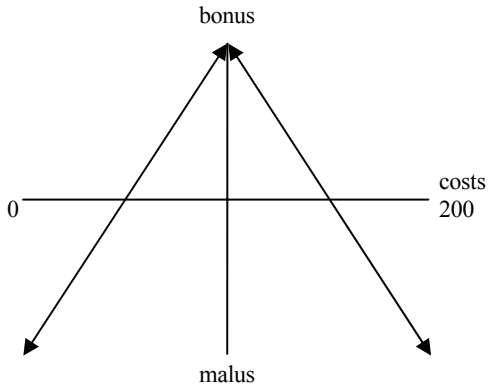
Figure 6.6. An incremental bonus system



A variant of the system in figure 6.5 is a mix between an inversely proportional system with target and a fixed bonus/malus system. It is characterized by an incremental bonus (figure 6.6). In the sense that there are – in this example – only two bonus levels, it is a clearer and (administratively) simpler version. Moreover, it agrees with the solution for the random variation in costs mentioned in the previous subsection: the definition of a range as norm. The physician may face an incentive just to cross the border between minimum and maximum bonus, but this holds for all break points. Moreover, reaching the maximum bonus is exactly what the physician is supposed to do.

Another variant of the inversely proportional system with target is displayed in figure 6.7. Here it is combined with a malus system through which two break points are created. Obviously, the incentives the physician faces are stronger.

Figure 6.7. Bonus/malus system with target



In case of a bonus or a malus system there is solely a settlement of surpluses or deficits at the end of a certain time period, resulting in a bonus or a malus or in an adjustment of the

payments in the next time period. A common variant of the malus system is the *withhold* in which case the third party anticipates a deficit by collecting a certain percentage of the fees paid to the physician. If the difference between the normative and the factual level of volume or costs is negative, then the withheld money covers the deficits. If positive, the withheld money is yet paid to the physician.

If a bonus or malus is based on the difference between the normative and the actual level of costs, the responsibility of a GP – and thus the derived bonus or malus – can be calculated by taking a percentage from the difference. As the risk is *shared*, the maximum percentage will be less than hundred percent, in absence of additional risk-reducing measures. If the physician's risk is limited, then the initial responsibility may be hundred percent though. Further, different percentages may be used for bonus and malus. The proportion will be determined by the amount of risk the third party wants to transfer, on the ability of the physician to influence volume or costs, and on the organisational and financial arrangements that are further made, like the presence of risk-limiting measures. The arrangements may also be refined by varying the percentages per type of care. Up to a fixed amount, the physician may be held fully responsible for a deficit in the primary care budget, but only partially for a deficit in the hospital budget. Equally, a surplus in the primary care budget may be returned to the physician. A surplus in the hospital budget, however, may be distributed among GPs, medical specialists and the third party. The latter may also have a right to a portion of the surplus, especially if it results from its managed-care activities. Another reason to limit the responsibility of the physician to only a proportion of the difference between normative and actual level is to limit the financial incentives he faces. In case of full budgeting or full capitation, the physician is fully responsible (i.e. bonus and malus amount to hundred percent). Such a large proportion, however, may result in cream skimming, in cost shifting, or in withholding or postponing care.

A bonus system may be combined with a malus system. A third party may use a malus as well as a bonus for the whole risk package. Another option is to use solely a bonus system for the hospital budget, but to use a bonus and a malus in the other budgets. In that case the physician is thus rewarded for his contribution to cost savings, but is not held responsible for deficits in the hospital budget. The use of only a malus system is unattractive for physicians but it may prevent increasing costs or sustain a certain level of care.

In three-tiered systems, the middle tier may change the arrangements made by the third party. It may introduce or abolish bonus, malus and withhold systems, or change the amount the physician is at risk. However, the middle tier will at least be motivated to 'pass through' the incentives it faces (Eggleston 2005). If a middle tier is made fully responsible for a certain risk package, it may share this risk with the physicians then. A middle tier sharing risk with the third party may also share its part with physicians.

6.4.6 Limitation of the physician's risk

Since the insurance risk and the risk of imperfect agency are difficult to separate, risk-sharing arrangements provide the physician with at least a portion of the insurance risk.

As a result, physicians made (partially) responsible for this risk perform a part of the insurance function and become, to some extent, insurer. But the goal of a financial-risk sharing arrangement is to stimulate a GP to act in the interest of the insured, and not to transform him into an insurer. For several reasons, providing a physician with much risk poses a hazard. Firstly, the physician may be prompted to take undesirable measures, like quality skimping, unnecessary referrals or cream skimming. Secondly, the incentive system may be ruined: a few expensive patients at the beginning of a financial period may dilute the incentives emanating from, for instance, a bonus system by making it virtually impossible for the physician to benefit from a surplus at the end of that period. Ultimately, a physician may fail to bear the risk and may be ruined.

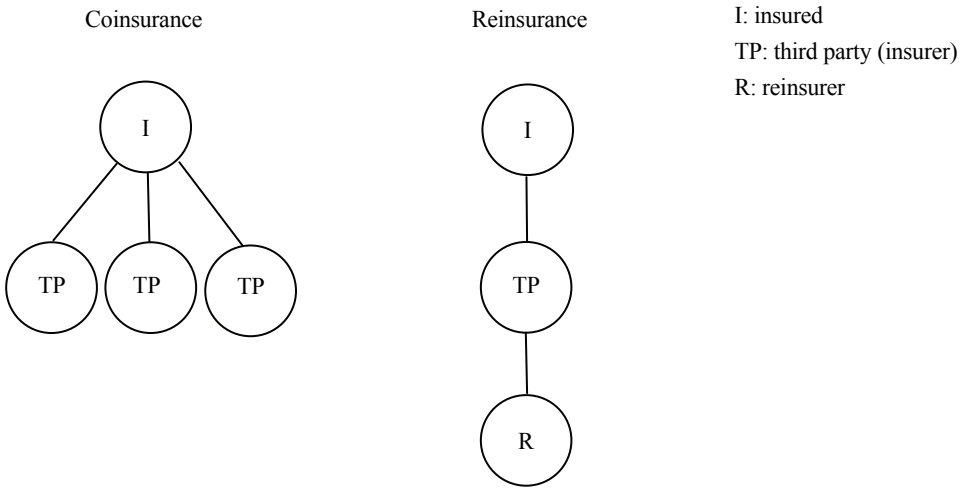
There are different approaches to limiting the risk a physician runs. The amount of risk is the result of the likelihood that the physician's patients need care (the probability of a loss), the type of care included in the risk package (the average size of the loss, i.e. low-cost versus high-cost care), and the amount the physician is liable for (like 10, 25, or 50 percent of the deficits). Evidently, a physician who bears full responsibility for all hospital care is subjected to a larger risk than one who is responsible for only 25 percent of primary care costs. A first approach is thus to limit the percentages and to design the risk package carefully. But even with relatively low percentages and relatively low-cost care, mere chance may significantly influence the likeliness a physician receives a bonus. Therefore, additional measures may be needed. Another approach is not to try to reduce the losses in advance, but to enhance the physician's ability to carry the financial consequences. For both approaches it would be useful to draw a parallel with strategies used by insurers to limit the risk, because the insurance industry has a long experience in this area. Hereafter, we consider the risk-spreading techniques that are applied in the insurance industry.

6.4.6.1 Risk-spreading techniques in the insurance industry

Two ways of risk spreading discerned within the insurance industry are coinsurance and reinsurance (Riley 1997).[15] In case of coinsurance, several third parties (insurers) accept a share of the risk and are only liable for the same share of the losses. Characteristic is thus that the insured concludes contracts with several insurers.

[15] The term coinsurance is often used in another context, namely to denote a specific form of cost sharing between insurer and insured. In such a coinsurance arrangement, claim costs are shared between insurer and insured in a predetermined rate. The idea in both contexts is the same: several parties accept a share of the risk.

Figure 6.8. Coinsurance versus reinsurance



In case of reinsurance, the insured has a contract with just one insurer. This insurer, on his turn, may conclude one or more reinsurance contracts with a reinsurer in order to spread his risk. Reinsurance may be defined as 'the insurance of contractual liabilities incurred under contracts of direct insurance or reinsurance' (Carter 1979, p. 4).[16]

Von Eije (1989) quoted Gerathewohl (1976) in discerning three types of technical dangers against which reinsurance may protect an insurer. One danger that may occur is the *danger of chance*. Usually, insurance premiums are based on independent claims, but sometimes, like in case of natural disasters, this assumption will not hold. The result may be that claims accumulate in a short period, which is thought of as being risky. Further, the danger of chance includes the random occurrence of a (large) number of independent (large) claims. Whether claims are large is considered in relation to premium income and reserves (Carter 1979, p. 7). Another technical danger is the *danger of being mistaken* in, for instance, rating or underwriting. Information problems, like adverse selection and moral hazard, may underlie this danger. Especially very large claims are considered to be dangerous then. A third danger is the *danger of change*. The insurer may be confronted with changes in factors that may influence (i.e. unexpectedly increase) the frequency and the average amount of the claims.

[16] The inclusion of the terms 'or reinsurance' in the definition of reinsurance can be justified as follows. In the first instance, a reinsurance contract between an insurer and a reinsurer provides insurance of liabilities that may be incurred under a contract between an insured and an insurer. In case the reinsurer, on his turn, has reinsured the reinsurance contract he has accepted – the reinsurance of a reinsurance contract is known as a retrocession – then reinsurance provides insurance of liabilities which may be incurred under a contract between an insurer and a reinsurer (Carter 1979).

Several forms of reinsurance can be distinguished.[17] A first form of reinsurance is the *quota share*. A quota-share arrangement is a proportional form of reinsurance in which a fixed proportion, for example 40 percent, of each claim is for account of the reinsurer.[18] Its administrative simplicity is a main advantage. It is, therefore, an inexpensive form of reinsurance. Moreover, the insurer is able to insure larger risks than would be possible without a quota-share contract. As the absolute variation in the retained share of the losses decreases, the probability of ruin – this is the probability that the insurer's losses will exceed its resources – will also decrease. Main disadvantages are that it also reinsures against small risks, and that it does not alter the relative variability of the expected losses on the retained share of the portfolio. Therefore, the loss ratio will remain the same (Carter 1979).

A second form of reinsurance is the *surplus* contract. This proportional reinsurance form is characterised by 'multiples of lines'. A line is the retention, i.e. the amount of a policy that the primary insurer retains under a surplus contract. By multiplying the line by a number – usually is agreed upon a number between five and twenty – the maximum liability for the reinsurer is found. The rest of the claim above the maximum is beyond the responsibility of the reinsurer, although a second surplus arrangement may be made to provide coverage for the costs beyond the reinsurer's initial responsibility. An important advantage is that, although depending on the amount of the line, small and medium claims are the insurer's responsibility, whereas large claims are partly covered by the surplus contract. Furthermore, it reduces the relative variability of the expected losses. The administration costs are a main disadvantage. For each policy has to be decided whether or not to reinsure, and for each claim has to be checked whether it should be covered by the reinsurer. In principle, it does not provide protection against large claims and against the accumulation of losses.

A third reinsurance method is the *excess of loss per risk* (i.e. per policy), which is a non-proportional form. Below a certain deductible, the claims are for account of the insurer. The amount above the deductible is the reinsurer's responsibility, although his liability may be limited up to a certain amount. An excess of loss per risk contract has several advantages. Firstly, it provides protection against large losses. Secondly, the insurer can retain a larger portion of the premiums, as he is liable himself for the more frequent losses below the deductible. Thirdly, it has the advantage of lower administration costs as, usually, deductibles are the same for all policies and premiums are a percentage of the sum of the primary premiums. However, it does not protect against the accumulation of claims due to a single occurrence or due to chance.

The fourth method of reinsurance, the *excess of loss per occurrence* (i.e. for all policies), is also an example of a non-proportional contract, but the goal is the reduction of the danger of chance by reducing the impact of dependency. This type of contract covers all claims with an amount above a certain deductible and that result from the same occur-

[17] Regarding the reinsurance methods heavily is drawn on the work of Carter (1979) and the thesis of Von Eije (1989).

[18] In proportional reinsurance arrangements the reinsurer is responsible for the same proportion of the claims as the proportion he receives of the premium.

rence, provided that the occurrence fits within the contract's definition. Deciding whether several claims originate from the same incident and whether this incident is indeed covered by the contract is a main problem of such arrangements.

A fifth reinsurance form is the non-proportional *stop loss*. This form reinsures the sum of the claim amounts as far as it exceeds a certain percentage of the sum of the premiums. It may protect the insurer against the risk of ruin. Often, however, the claim amounts above the retained percentage are shared between primary insurer and reinsurer. Moreover, the reinsurer's liability is often limited to a maximum amount. Some consider the stop loss to be an ideal form of reinsurance for the insurer as the cause of the total claim amount is irrelevant. It provides protection against increases in the frequency as well as in the size of losses.

A final form of reinsurance mentioned here is the *n largest claims*. In a simple form, the insurer reimburses the highest n claims in full. It is a non-proportional method. For the reinsurer this form has the advantage over the excess of loss per risk that the number of claims will not increase due to inflation.[19] Inflation may increase the size of the claims though. A solution may be to use a deductible that is revised in case of inflation.

6.4.6.2 Risk-spreading techniques in the health-insurance industry

The risk-spreading techniques described above are applied within the insurance industry in general, but are of use in the health-insurance market as well. Many health plans (HMOs for instance) have reinsurance arrangements to cover, at least partly, those claims that exceed a certain deductible. These deductibles usually relate to the claims of a single risk or to the aggregate claim costs – i.e. an excess of loss per risk or a stop loss respectively (see Kongstvedt 1993a, Brennfleck Pascuzzi 1993, and Ward 1993).

Van Barneveld et al. (1998) analysed whether different forms of mandatory reinsurance could be used as a supplement to risk-adjusted capitation payments within a health-care system with competing health insurers. Such techniques should reduce incentives a health insurer may face to take undesirable measures – especially cream skimming – but maintain incentives for efficient behaviour. They distinguished three reinsurance forms. In the first form, *high-risk pooling*, insurers are allowed to select a fraction of their portfolio of which the costs will be pooled (partially). A small modification of this form might be the replacement of a fraction by a fixed number of insured whose costs are pooled. Then, there is some resemblance to the n largest claims form mentioned above.[20] An important difference, however, is that with high-risk pooling insured are selected in advance of a certain financial period, whereas in a n largest claims system insured are selected at the end of the financial period. Their second form, the *excess of loss*, resem-

[19] In case of an excess of loss per risk inflation may increase the size of the loss beyond the deductible. Hence the number of claims for the reinsurer may increase as well.

[20] The terminology is somewhat confusing. Pooling resembles coinsurance (see subsection 6.4.6.1) but is used by Van Barneveld et al. in the context of reinsurance. The explanation for this is that coinsurance is presented here as a form of reinsurance: the insurer reinsures a fraction of his portfolio by pooling it with other insurers.

bles the excess of loss per risk form in the above. The third form, *proportional pooling*, is what is labelled here a quota-share arrangement.

Despite the similarities between the risk-pooling arrangements known from the insurance industry in general and those discussed by Van Barneveld et al. (1998) for the health-insurance industry, there are some differences. First of all, the goals of applying such arrangements are different. Goal of reinsurance techniques used within the insurance industry is mainly to offer protection against the random risk of large individual losses or of an accumulation of losses due to a single occurrence, and against fluctuations in the aggregate loss experience (Carter 1979). Van Barneveld et al. (1998) argued, however, that in health care a regulator's goal will be to reduce the incentives for cream skimming by reducing the predictable risk that may occur as a result of using capitation payments which are imperfectly adjusted for risk. They considered it the insurer's and not the regulator's job to deal with random fluctuations. Competition and premium regulation in a health-insurance market, however, may restrain the insurer's ability to adjust premiums to random fluctuations. It may very well be the case, therefore, that a voluntary reinsurance arrangement is used in conjunction with a mandatory pooling arrangement. A second difference is that, usually, premiums for reinsurance policies are adjusted for risk, whereas in the variants of Van Barneveld et al. the premiums are independent of the risks that are pooled in order to prevent cream skimming by the insurer. It should be noted that this pertains to the specific situation as described by Van Barneveld et al. in which there is only one reinsurer, i.e. no competitive reinsurance market. As is the case in the traditional reinsurance market, competition between reinsurers results in risk-rated premiums. A third difference is that, usually, reinsurance arrangements are voluntary. The variants discussed by Van Barneveld et al., however, are mandatory.

6.4.6.3 Reinsurance techniques in risk-sharing arrangements

The question is whether the goals as well as the ways of applying risk-spreading techniques within a system of risk-bearing GPs differ from those described above.[21] A first objective may be to reduce the incentives for cream skimming and also for cost shifting and quality skimping, i.e. to reduce the predictable risk. A second objective may be to protect the physician against the random risk as well. Due to the relatively small practice size, a GP is especially vulnerable to random fluctuations in occurring costs.

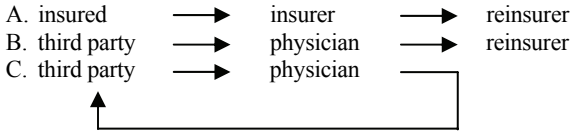
Reinsurance techniques may be applied in different ways. Regarding the way an arrangement is financed, reinsurance premiums should not be related to the risk of the patients whose costs will be pooled as otherwise the physician would still face an incentive to skim cream (in conformity with Van Barneveld et al. 1998, p. 225). The problem of premium rating is most noticeable in case a GP has to decide whether or not to take out reinsurance for an individual risk. However, the most common way to finance, for instance, an excess of loss per risk or a stop-loss arrangement is by means of a premium calculated as a percentage of the total premium income. For the physician, this would mean that reinsurance coverage is provided for all his patients against a premium that is a percentage of his gross income. This would reduce the danger of cream skimming con-

[21] See also the remarks on this subject by Van Barneveld et al. (1998, p. 231).

siderably. Still, it will depend on the reinsurance technique and its exact application whether there remains a retention, i.e. a quantity not covered by reinsurance, and thus whether there remains an incentive for undesirable behaviour. With regard to the voluntariness of risk-spreading arrangements, a third-party agent may decide to compel the contracted physicians to arrange for risk-spreading systems or to accept the arrangements made by the third party. This may be the best way to protect the incentive system and to guard the insured' interests.

Risk-spreading arrangements within a system of risk-bearing GPs may differ in another way from those in the insurance industry. A GP may seek reinsurance with a risk-rated premium from a party outside his relationship with the third-party agent (see figure 6.9 B), which resembles normal practice in the insurance industry (figure 6.9 A). However, another option is that the third party functions as reinsurer (figure 6.9 C). For example, the third party pays the physician a budget, but limits the physician's liability to amount *x*. Costs above this deductible are covered by the third party itself. Clearly, such a design results in a thin line between the basic payment system and the reinsurance system.

Figure 6.9. Risk-spreading arrangements



A system in which the third party functions as reinsurer has several advantages over a system in which the physician has to seek reinsurance from another party. Firstly, the administrative costs will be lower. In case costs arise that are covered by reinsurance, the physician only has to deal with the primary insurer, i.e. the third party. Further, the terms of a reinsurance contract will be related to the contract between third party and physician. So if a patient changes from third party, if a patient changes from physician, or if the terms of the contracts between third party and physician are changed, the physician does not have to change his reinsurance contract.

A second advantage is that such a system may be less expensive as there is only one insurer involved instead of two, who probably would both charge a loading fee for profit as well as a risk load.

A third advantage – and this is important from an agency point of view – is that reinsurance changes the incentives a physician faces. By offering reinsurance-like protection itself, the third party will be more able to control these incentives.

In case the third party is the government, reinsurance by the government has the major advantage that the costs of reinsurance can be spread very broadly. Financing of the reinsurance program can be spread over the whole population (Blumberg and Holahan 2004).

On the other hand, in one specific (but not exceptional) case there is an advantage in a system in which the physician himself has to seek reinsurance: if a physician has contracts with several third parties. If all individual third parties would base their measures to

reduce the physician's risk on the proportion of their insured in the physician's practice population, then the physician would be 'over-insured'. An option for the physician is then to seek reinsurance from another party or to pool the contributions of the different third parties with other physicians. This latter option, risk-pooling with other physicians, is not only a potential solution to the problem of several third parties. Pooling may be used next to a reinsurance technique, or as a fully-fledged alternative in order to spread the physicians' risks (see the next subsection).

In principle, the several reinsurance techniques described in subsection 6.4.6.1 can be applied to the practice of the GP. In a *quota-share arrangement*, the physician is responsible for x percent of the patient's costs. The remaining $(100 - x)$ percentage is for account of the third party (or a reinsurer). In a *surplus arrangement*, the third party covers the patient's costs up to x times the physician's retention. If an *excess of loss per risk* is used, an individual patient's costs above a certain deductible are for account of the third party. The *excess of loss per occurrence* form reduces or removes the physician's liability in case of a catastrophic incident covered by the policy, like an epidemic or a natural disaster. A *stop-loss contract* covers all the patients' costs a physician occurs as far as these costs exceed a particular percentage of the payments the physician received from the third party for the patients involved. In case of a *n largest claims,* finally, the costs of the n patients with the highest costs are the third party's liability.

With reinsurance forms for which this is not already made explicit by nature of the contract, the maximum liability of the third party can be fixed or may be limited up to a certain percentage of the patient's costs. In the latter case, a combination may be made between, for instance, an excess of loss per risk and a quota share. The rationale for such additional arrangements is not so much to protect the reinsurer from too much risk – as is the case in the regular reinsurance contracts in the insurance industry – but to preserve an incentive to provide cost-effective care.

Other additional arrangements are to make the retention (deductible) depending upon the number of patients covered (the more patients, the higher the retention), or the type of care (a lower retention for high-risk care).

The above mentioned techniques as well as the risk-pooling techniques discussed in the following subsection can be applied to reduce a physician's risk. An alternative approach might be to make the physician bearing the risk, and to compensate him by means of a risk load. This load can be used to create a buffer (a reserve) that will help the physician to meet cost fluctuations (Tolley et al. 1987).

6.4.6.4 Risk-pooling techniques in risk-sharing arrangements

Risk pooling resembles the aforementioned technique of coinsurance (see subsection 6.4.6.1). The risk is not transferred to a single GP but spread over a group of physicians. All members of the pool take a share of the risk the third party wants to transfer to the physicians. A risk pool may thus be defined as a group of providers who share in the rewards and penalties from surpluses and deficits in the budgets for certain types of care (Hillman et al. 1992). These providers may be GPs but, as will be argued in the following text, other providers may be members as well. Within the pool, the physicians mutually

share the risks and thus the bonus or malus that may result from the surpluses or losses in the pooled budgets. It is a means to reduce the risk an individual physician runs.

In determining the size of a risk pool, spreading of risks needs to be balanced against financial incentives. Pool size is negatively related to risk but also to the incentives an individual pool member faces to behave cost-effectively. If the size of the pool increases, the effect of the physician's behaviour decreases as the costs are borne by the group. Hillman et al. (1992) found that in 1988, of 216 three-tiered HMOs about 15 percent had 'risk pools' that consisted of individual primary care physicians. About 30 percent had risk pools of physician subgroups, averaging 37 physicians. About 38 percent of the HMOs grouped all their primary care physicians in a single risk pool with on average 275 physicians (data on the remaining HMOs were missing). The physicians in the individual 'pools' were 100 percent responsible for the risk. Physicians in the subgroups as mentioned were responsible for, on average, 2.7 percent (1/37), whereas their colleagues in the single risk pool were responsible for only 0.4 percent (1/275) of surpluses or deficits.

It is not only the number of providers in a risk pool that is important for the size of the risk. Equally important are the number of patients per physician (this determines the individual physician's share) and the total number of patients for whom money is put into the pool (the risk per physician decreases with increasing size). According to the law of large numbers, the standard deviation of the loss per patient will reduce if the number of patients increases.

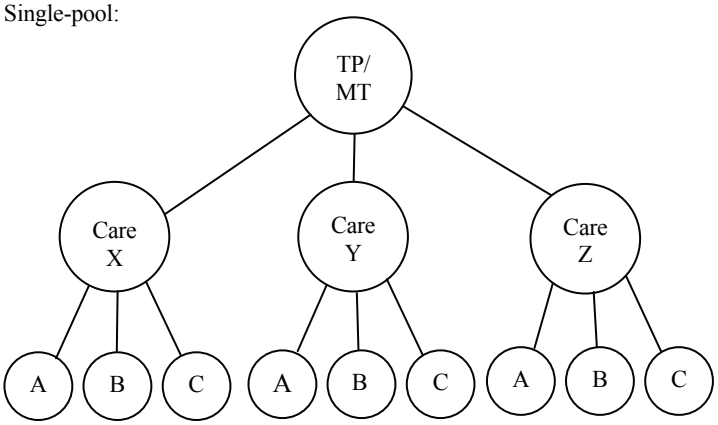
Besides the size of the pool, the proximity of its members may also be of influence on the incentives an individual member faces. A risk pool consisting of a group of GPs practising in the same building will probably have a larger effect than a risk pool of which its members are practicing individually within a wide area (Hillman et al. 1992). Finnish experiences, on the other hand, showed that working within the same organisation is more important than working in the same building (Van de Ven and De Jong 1992).

Another point, mentioned by Hillman et al. (1992, p. 139) is the culture of the risk pool. This culture will result from the amount at risk, the size of the pool, and the location of its members, but also from the 'philosophy of the risk pool'. Some risk pools seem to be more responsive to managed-care techniques employed by the third-party agent or the pool itself than other pools.

A GP may be member of one pool as well as member of several pools (a multi-pool system). In the first case, the members of the pool are (partly) responsible for the difference between the normative and the actual level of volume or costs of all the care included in the risk package. In the second case, the physician is member of several risk pools. In a multi-pool system, the pools have, in principle, different members (see figure 6.10).

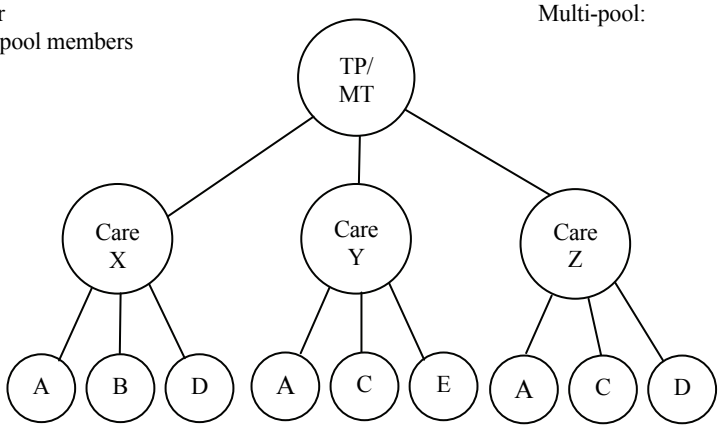
Figure 6.10. Single-pool versus multi-pool system

Single-pool:



TP: third party
MT: middle tier
A: general practitioner
B, C, D, E: other risk-pool members

Multi-pool:



Creating several risk pools has the advantage that surpluses and deficits can be distributed among the members in various proportions. GPs may be held fully responsible for the differences between the normative and the actual level of volume or costs within the primary care fund, but may be held responsible for, for instance, only twenty percent of such differences within the hospital risk pool. Related is the advantage that the composition of a pool can be made contingent upon the type of care and upon the extent of the providers' influence on the final results. The primary care risk pool may consist solely of GPs, whereas the responsibility for the hospital risk pool may be shared with medical specialists and the hospital itself. The risk pools may vary in size dependent on the risk associated with the type of care. Hospital care will demand a larger pool size, other things

being equal, than primary care. A disadvantage of creating several risk pools is that incentives for substitution of care may be decreased. Another disadvantage is the increased complexity of the arrangements.

Risk-pool arrangements may be two-tiered as well as three-tiered. In a two-tiered system the third party settles the surpluses or the deficits between the members of the pool. In a three-tiered arrangement the middle tier receives payments from the third party and then settles the surpluses or the deficits.

6.5 The third party's options for dealing with risks

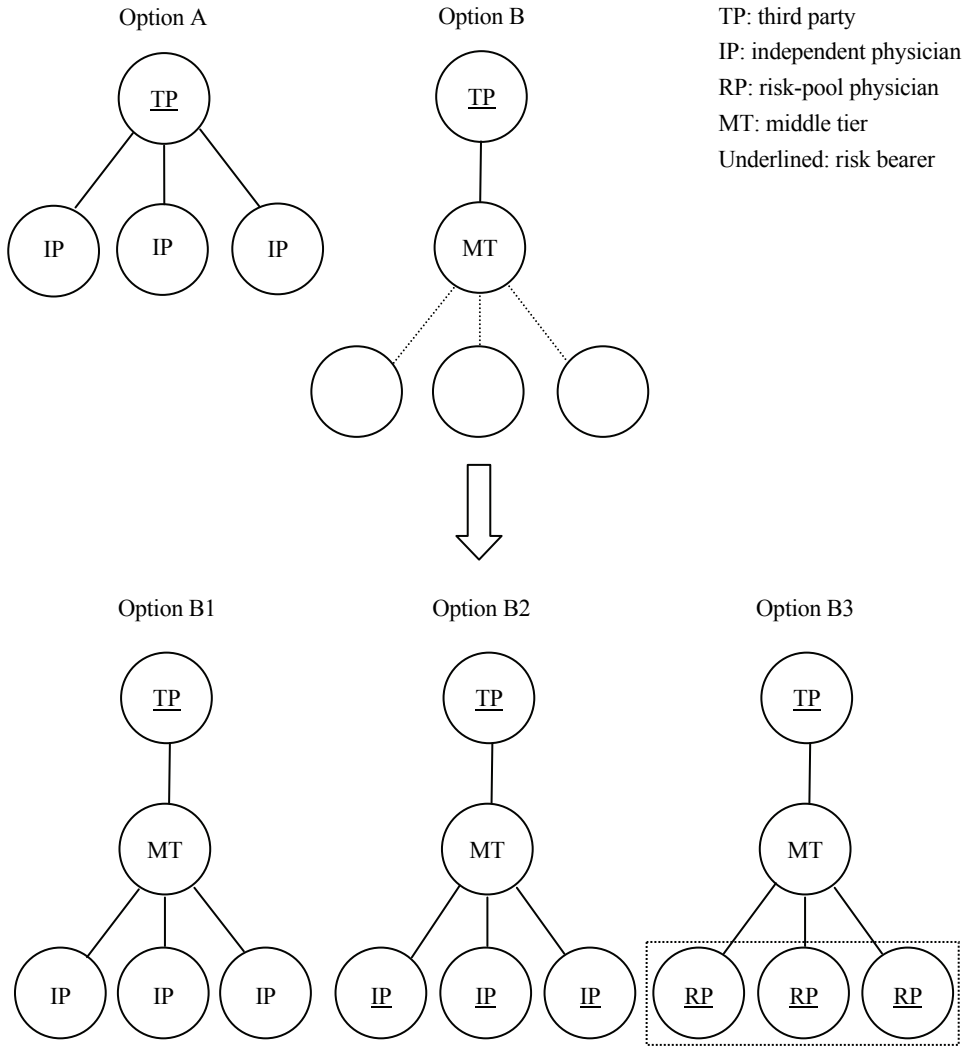
As argued in subsection 6.2.4, the third party has three feasible options to deal with the financial risks: risk bearing (hundred percent responsibility for the third party), risk shifting (zero percent responsibility), and risk sharing (responsibility larger than zero but smaller than hundred percent).[22] Formation of risk pools creates additional options in case of the risk-shifting or risk-sharing strategy. Except shifting the risk to or sharing it with independent GPs or middle tiers, the third party may also shift it to or share it with a risk pool of GPs. The three strategies a third party may employ in combination with the strategies a middle tier may employ result in a set of options for the allocation of the risk. Gold et al. (2002) pointed at the fact that complex arrangements can even lead to four, five or even more tiers (multi-tiered arrangements). This can make it hard to identify the location of the risk.

The first strategy for a third party is to *bear the risk* itself (denoted by bold letters in figure 6.11). In this case, there are two main options. The first option (A) is to bear the risk and pay independent GPs according to, for instance, a fee-for-service system. Option B is to pay an intermediary organisation (the middle tier) likewise.[23] The middle tier, then, has three options: paying independent physicians according to a fee-for-service system (option B1); paying independent physicians according to a risk contract with, for instance, capitation payments (option B2); or paying a group of physicians (the risk pool) according to a risk contract (option B3). Options B2 and B3 – in these options the third party thus pays the middle tier according to a risk-free contract and allows the middle tier to pay the physicians according to risk contracts – are unlikely and, therefore, mainly of theoretical importance.

[22] The fourth option, risk splitting, is left out of consideration here. As argued, the middle coarse in which the third party bears the insurance risk and shifts the risk of imperfect agency to GPs is practically unfeasible.

[23] Note that the option of a third party paying a group of risk-pooling GPs directly (comparable with options D and G in figure 6.13 and 6.14 respectively) is omitted here. Since the third party bears all the risk, creating a risk pool makes no sense.

Figure 6.11. Risk bearing by the third party



A second strategy for a third party is to *shift the risk* (see figure 6.12). Now there are three main options. The first option (C) is to shift the risk to independent physicians by paying them, for instance, full capitation. A third party may also choose to contract a risk pool of physicians according to a risk contract, which is a second option (D). A final option is to contract a middle tier (E). The middle tier, on its turn, may choose one out of five options. These also boil down to bearing, shifting and sharing. The middle tier itself can bear the risk (E1) or shift it to independent GPs (E2) or to a risk pool (E3). In options E4 and E5, finally, the risk is shared with independent physicians or with a pool of physicians.

Figure 6.12. Risk shifting by the third party

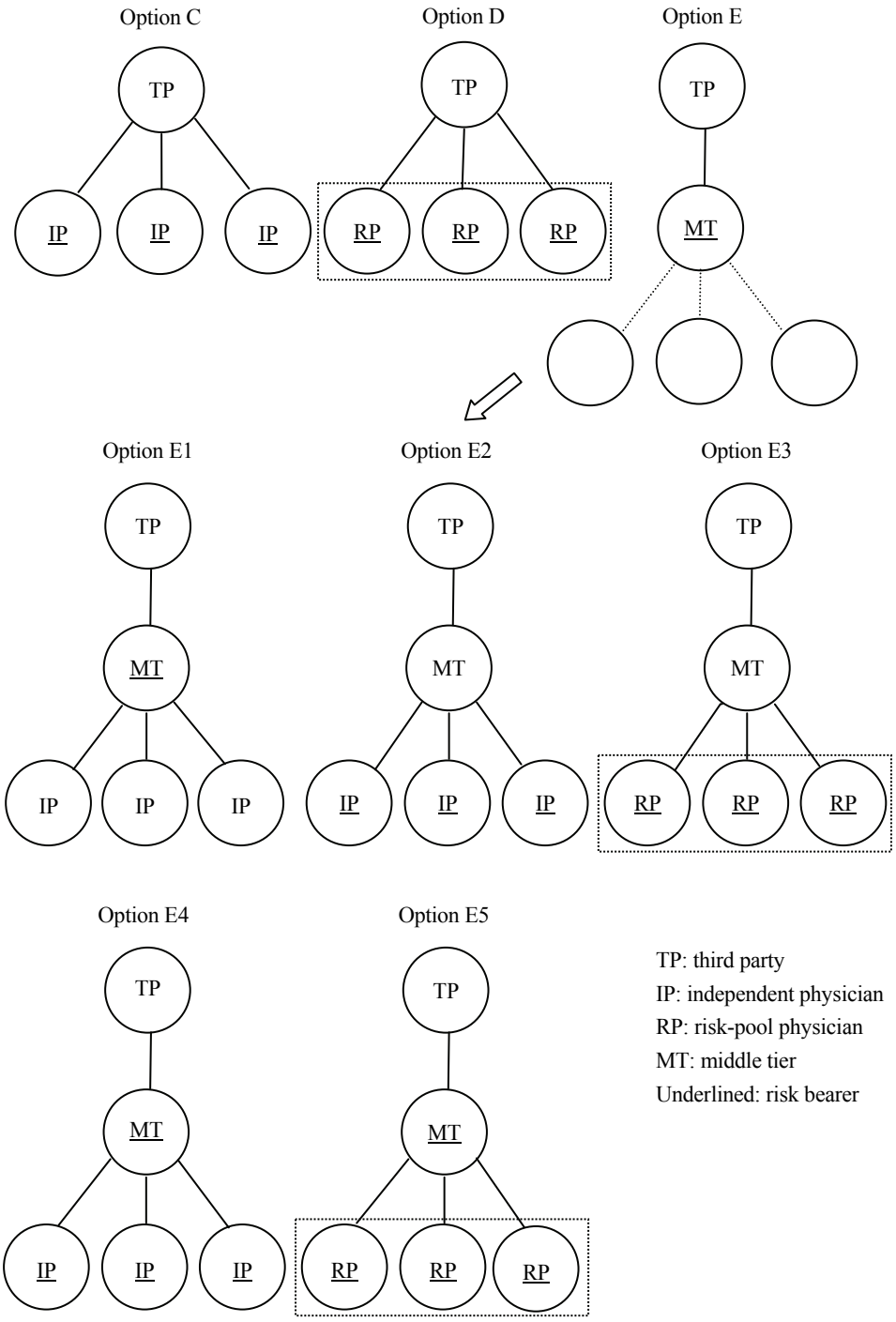
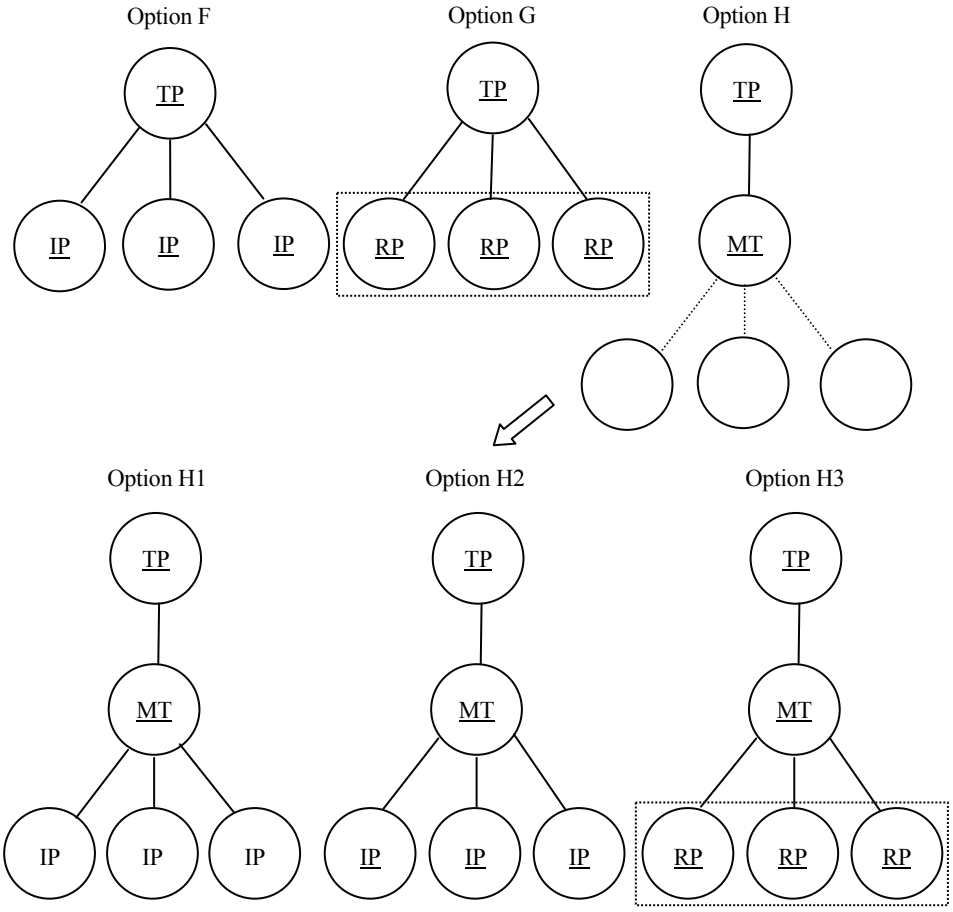


Figure 6.13. Risk sharing by the third party



TP: third party
IP: independent physician
RP: risk-pool physician
MT: middle tier
Underlined: risk bearer

The third strategy for a third party is to *share the risk* (see figure 6.13). A first possibility then is to share the risk with independent physicians (F). Another option is risk sharing with a risk pool of GPs (G). Option H contains of risk sharing with a middle tier which, on its turn, may bear its part of the risk (H1), or may share it with independent GPs (H2) or with a pool of physicians (H3).

In the options B, E, and H, a middle tier subcontracts independent physicians or pools of physicians. It may conclude the same type of contracts with these physicians as with the third party, or it may alter the nature of the contracts by changing a risk contract in a risk-free contract. Whether a third party exerts influence on the contracting of physicians by middle tiers may differ. A third-party agent will probably prohibit a middle tier from shifting all the risk to physicians as this may create conflicting interests, but a third-party payer may not. Hillman et al. (1992, p. 138) noted that third parties – HMOs in their research – are sometimes ignorant of the way a middle tier contracts physicians.

In option E the middle tier has the disposal of the largest set of contracting modalities, namely five. Not only has it the possibility to bear all the risk or to share it with physicians, it also has the option to shift the whole of it to physicians. The other options with a middle tier, options B and H, have the advantage that the third party retains some influence by restricting the possibilities for the middle tier, and that the physicians are prevented from bearing all the risk.

Table 6.2. Risk-bearing party per option

Risk bearer	A	B1	B2	B3	C	D	E1	E2	E3	**E4**	**E5**	**F**	**G**	H1	**H2**	**H3**
TP	X	X														
IP					X			X								
RP						X			X							
MT							X									
TP + IP			X									X				
TP + RP				X									X			
TP + MT														X		
MT + IP										X						
MT + RP											X					
TP + MT + IP															X	
TP + MT + RP																X

TP: third party
IP: independent physician
RP: risk-pool physicians
MT: middle tier
Bold: Physician risk-sharing options

In table 6.2 the allocation of the risk is represented in a different way. It can easily be seen now that in the options C, D, E2, and E3, the physicians bear all the risk. In the options B2, B3, E4, E5, F, G, H2, and H3, the physicians are partly responsible (risk-sharing). Actually, in the unlikely options B2 and B3 the risk is not really shared and, therefore, these are further left out of consideration. Of the sixteen options considered in this subsection, only six remain as options in which the risk is shared with physicians

(printed bold in the table). A third-party agent wanting to control GPs by means of risk sharing will thus have to:
- share the risk directly with physicians (options F and G),
- shift the risk to a middle tier and arrange that the middle tier shares it with the physicians (options E4 and E5), or
- will have to share it with the middle tier and then make such arrangements (options H2 and H3).

6.6 Summary and conclusion

One of the functions of a third-party agent is to provide insurance against the *insurance risk*. This risk results from the occurrence of illness, which has largely a stochastic nature. In return for an insurance premium, the risk is transferred from the insured to the third party. The risk transfer involves a second risk, though, which we labelled the *risk of imperfect agency*. It consists of the risk of the provision of cost-ineffective care (mainly resulting from over-provision and inappropriate care) and of the risk of underprovision of care. The risk of imperfect agency can result from agency problems within the relationship between patient and physician, and from the presence of insurance. Health insurance may lead to consumer-induced moral hazard and to supplier-induced moral hazard.

As the provision of health care is to a large extent at the GP's discretion, the third-party agent may focus on this physician in order to reduce the size of the risk of imperfect agency. Once taken over the insurance risk and the risk of imperfect agency, the third-party agent may choose between four strategies in order to handle them. The first option is *risk bearing*, in which the third party accepts the risk of imperfect agency as well as the insurance risk. This option is not compatible with the agency function because the third party makes no attempt to influence the way health care is delivered. The second option is *risk shifting*, in which the third party shifts both risks to individual (primary care) physicians or to a group of (primary care) physicians. The physicians become responsible for the insurance risk, which however is typically a third-party function. Moreover, the responsibility for both risks may prompt the physicians to take, from the third-party agent's viewpoint, undesirable measures to reduce their risk, like cream skimming. Hence this option is not compatible with the agency function either. In the third option, *risk splitting*, the third party attempts to separate both risks. Theoretically, this middle coarse between risk bearing and risk shifting is the most satisfactory solution. The third party can shift the risk of imperfect agency to the physicians whilst retaining the insurance risk itself. In practice, however, it will be very hard to separate both risks. The final option is *risk sharing*, which also has the advantage of being a middle course between risk bearing and risk shifting but with the side-effect of shifting a part of the insurance risk to the (group of) physicians. The difficulty of separating both risks is evaded though.

The answer to the first research question of this chapter,

What is the rationale for financial-risk sharing between third-party agents and general practitioners?

consists thus of two parts. Firstly, the cost-effectiveness and the quality of care are to a large extent at the discretion of GPs. Hence for a third-party agent these physicians are a perfect starting point for improvements in the provision of care. Secondly, of the four strategies to handle the risk, risk sharing is the best for a third-party agent. The other three strategies are inconsistent with the third party's agency function (risk bearing and risk shifting) or not very feasible (risk splitting).

In a financial-risk sharing arrangement the third-party agent stimulates the GP to reduce the amount of cost-ineffective care. Because the risk is shared, the incentives for over-provision and for underprovision can be balanced. The effectiveness of such an arrangement, however, depends on its financial and organisational structures. The more the arrangement has the effect of the risk-shifting option, the stronger are both the incentives for cost-effective behaviour and for undesirable physician behaviour. There are at least three obvious forms of undesirable behaviour that a GP may show if he is at risk. In case of cream skimming, the physician selects patients for whom the (expected) costs are lower than the (expected) reimbursement. In case of cost shifting, the physician provides or arranges care for which he is not financially responsible. In case of quality skimping, the physician postpones or even withholds care, reduces his efforts, et cetera.

The second research question of this chapter was:

How can systems of financial-risk sharing be structured?

Financial-risk sharing arrangements have five main aspects that the third-party agent has to take into consideration while drawing up the arrangement. These aspects are the risk package, the size of the practice population, the normative level of care, the bonus system, and the limitation of the physician's risk.

A first crucial aspect is the scope of the health-care goods and services for which the physician is financially responsible, the *risk package*. The type of care included will determine the probability that the physician incurs costs as well as the variability of these costs. For some types of care it will be easier to diagnose and to estimate the costs of treatment, which is especially important if the GP has to arrange and pay for follow-up care. Other matters to take into consideration are whether the risk package provokes cream skimming or cost shifting, whether it is divided into separate cost categories and whether the behaviour of other providers of care is influenced.

The second aspect is the *size of the practice population* or the proportion of it for which the physician is financially responsible. The relative and absolute size of this population determine the magnitude of the incentives, the ability of the physician to shift costs to (patients from) other third parties and the extent to which the physician is vulnerable to random fluctuations in the costs.

A third aspect is the *normative level of care*. The third party may define a norm, for instance in terms of a certain volume of care or a cost level, with which (the outcome of) the physician's behaviour is compared. This is probably the most difficult part of the arrangement. It is hard to determine an optimal level based on medically necessary, cost-

effective and needs-based care. A way to determine a norm, then, is to use actual costs and to base the norm on historical costs or, better, on average costs. The norm can be adjusted for systematic differences in health status and for some of the other systematic factors, in so far as the physician can not influence them.

The fourth aspect is the *bonus system*. Eventually, the physician's financial responsibility may find expression in a bonus, which is an ancillary payment made if the physician has met certain requirements (like a financial norm). The negative variant of it is the malus, which the physician has to pay if he has not met the requirement. A bonus may be a fixed amount or may be proportional or inversely proportional to the difference between actual and normative costs. Further, the norm may function as a threshold or as a target. In the latter case, the bonus is at its maximum once the norm is met, which will probably be the third-party agent's first choice.

The fifth aspect is the *limitation of the physician's risk* by means of additional measures. These measures are considered additional because the amount of risk is in the first instance the result of the risk package, the practice population, et cetera. The difficulty of separating the insurance risk from the risk of imperfect agency makes that the physician is also responsible for (a part of) the insurance risk. Without limiting this risk, there is a chance that the physician shows undesirable behaviour, that the incentive system will malfunction or that the physician is ruined. The incentive system may not function properly due to a few expensive patients in the first part of the financial year.

One way of risk reduction is reinsurance. A GP who has taken over part of the third party's risk may on his turn insure his liabilities. Although in the regular insurance industry reinsurance usually results in a risk contract with a second insurer, in the present risk-sharing arrangements the third-party agent may function as a kind of reinsurer. Reinsurance is then a part of the financial arrangement between third party and physician. Not only may this result in lower costs, it also has the advantage that the third-party agent may balance the incentives from the reimbursement system with the incentives from the reinsurance system. Examples of reinsurance systems are an 'excess of loss per risk', 'an excess of loss per occurrence', a 'stop loss' or a 'n largest claims'.

Another way to reduce the physician's risk is by means of risk pooling, which is a form of coinsurance. In a risk-pooling arrangement a group of GPs share together, possibly with other providers of care, in the rewards and penalties from surpluses and deficits in the budget(s) for a defined health-care package. Several variables determine the incentives emanating from the risk-pool arrangement, like the number of physicians or other providers, the number of patients, the proximity of the members of the pool, et cetera. Other ways to vary the arrangements are by creating a multi-pool system and by adding intermediate organisations: the so-called middle tiers. This results in a myriad of options to allocate the financial risk, but only in a limited number of options the risk is really shared with physicians. A third party that wants to control physicians by means of risk sharing has to share the risk directly with the physicians, or has to arrange that the middle tier shares the risk with the physicians.

7 FINANCIAL-RISK SHARING IN PRACTICE

7.1 Introduction

As each GP receives payments, and as incentives emanate from all payment systems, each GP faces a financial incentive. Not each GP faces a financial risk (as defined in subsection 6.2.1), however. Whether a physician faces such a risk, depends on the basic payment system as well as on the presence and nature of an ancillary payment system. In some health care systems GPs are being put at risk for follow-up costs, like costs of drugs or hospital care. This is no new phenomenon. Already in the first half of the twentieth century there were third parties and physicians that agreed on risk contracts. Examples are the 'Zaanland system' and the 'Amsterdam system' in the Netherlands and the first Prepaid Group Practices in the United States.

In the previous chapter, we discussed the rationale and the structure of arrangements in which the financial risk for follow-up costs is shared between third parties and GPs. By dividing the arrangements into five distinct aspects, an analytical framework has been created consisting of 1) the risk package, 2) the size of the practice population, 3) the norm, 4) the bonus, malus or withhold system and 5) the risk-limiting measures. We argued that these aspects can be chosen in various ways, with numerous arrangements as a result.

An analytical framework of risk sharing serves two purposes. Firstly, it can be used to structure new arrangements. Secondly, such a framework can be used to analyse and to evaluate risk-sharing arrangements in practice. The requirement then, is that it sufficiently provides insight into the key differences of various (formerly) existing arrangements. If so, one may use these insights to infer from evidence of specific arrangements the effectiveness of similar arrangements in stimulating the GP to act as agent for the patient. This second aim of the framework is pursued in the present chapter. Hence an answer is sought to the following two research questions:

- *What are actual effects of different systems of financial-risk sharing on the performance of general practitioners?*

- *Does the analytical framework of financial-risk sharing sufficiently provide insight into the key differences of systems in which the risk is shared between third party and general practitioners so as to infer the effectiveness of such systems?*

In the following subsections, several older as well as more recent experiences with risk sharing will pass in review. Chosen is for examples from the Netherlands, the United Kingdom and the United States. The rationale for this choice is that third parties and particularly GPs are clearly present within the health-care systems of these countries. Moreover, these countries have (some) experience with third parties that are attempting or

have attempted to exert influence on the provision of health care by GPs, or who have shifted part of the third-party functions to these physicians. Especially experience with the use of financial incentives in general and risk-sharing arrangements in particular was a decisive criterion for our selection of examples.

7.2 Risk-sharing experiences in the Netherlands

7.2.1 Introduction

GPs have an important position within the Dutch health care system. As opposed to medical specialists, they are generalists and directly accessible for their patients. Further, they act as gatekeepers; patients have to be referred by their GP otherwise the health insurer does not pay for health-care costs. There is usually a direct relationship between the GP and the health insurer (a contract model and a two-tiered system). Only some GPs are employed by a middle tier; either another GP or a primary health centre. The contract between health insurer and GP specifies administrative, quality and financial aspects of their relationship.

Of the three countries described, the Netherlands has the least experience with third parties and GPs sharing financial risks for follow-up costs. Until 2005, sickness funds paid GPs mainly by capitation. Private health insurers had no contractual relationships with GPs, but reimbursed the health care expenses of the insured (reimbursement model). Private patients paid their GP per consultation. In principle, Dutch GPs bore the risk of deficits in their capitated primary-care budgets for their public patients (about two third of their practice population). In practice, however, they were able to shift costs to sickness funds or to private health insurers – by prescribing drugs or referring patients to other providers, or by increasing the revenues for private patients. Hence they only partially bore (i.e. shared) this risk.

Since decades there has been a lively debate on the 'ideal' compensation system for GPs (see Vermaas 1994 for a short review). More recent proposals for payment reforms also included financial-risk sharing for follow-up costs. In 1993, the 'Ziekenfondsraad' (Council of Sickness funds) proposed a salary or a capitation system for public as well as private patients in combination with bonuses for efficient prescribing and referral behaviour (Ziekenfondsraad 1993). In 1994, the government appointed Commission on Modernising Curative Care proposed to replace the existing systems with a capitation system adjusted for age (eighty percent of the compensation) combined with locally differentiated bonuses for the performance of specific functions and bonuses for efficient behaviour (Commissie modernisering curatieve zorg 1994). Also in 1994, researchers, a health insurer and GPs developed a model for a contract between health insurer and GP that contained an arrangement of financial-risk sharing (Breedveld et al. 1994). In 1995, the associations of private health insurers and GPs expressed their intention to replace the fee-per-consultation system with a capitation system. Even risk-sharing arrangements, like a fundholding experiment or a bonus system were suggested (Paritaire Werkgroep Huisartsenzorg 1995). In the same year the employers association proposed financial

incentives for GPs as one of the possible ways to influence their referral and prescribing behaviour (VNO-NCW 1995).

In 1999, four reports were published in which risk sharing in some form was proposed. In three of them, a bonus system was proposed to stimulate GPs to prescribe drugs efficiently (Begeleidingscommissie Uitvoering Geneesmiddelenbeleid 1999, The Boston Consulting Group 1999, MDW-werkgroep Geneesmiddelen1999). In the fourth report, a budget system for GPs was proposed (Max Geldens Stichting 1999).

In 2001, the government appointed Commission on the Future Financing Structure of Primary Care, after its chairman called the 'Tabaksblat Commission', proposed a radical redesign of the payment system. The two existing payment systems should be replaced by a uniform system (i.e. no difference between public and private systems) with separate payments for practice costs and income. The basic payment system contained several incentives for efficient behaviour. In addition, the committee proposed the use of ancillary payments for adequate referral behaviour and for adequate prescribing behaviour (Commissie Toekomstige Financieringsstructuur Huisartsenzorg 2001). In 2004 the Minister of Health (Minister van Volksgezondheid, Welzijn en Sport 2004) and the College for Health Care Tariffs (College Tarieven Gezondheidszorg 2004) both proposed a revision of the payment system, partially in line with the proposals of the 'Tabaksblat Commission'. It was not until the middle of 2005, however, before a compromise agreement was reached between the National Association for General Practitioners (LHV), the Ministry of Health (VWS) and the Dutch Association for Health Insurers (ZN). From 2006 GPs receive a partial capitation per patient per year plus fees per consultation. At least two remarkable arrangements are made. One arrangement concerns a direct relationship between the performances and the financing of the profession. Savings as a result of efficient behaviour are invested in GP care. Another remarkable arrangement is that, at least in 2006, the financial risk associated with the volume of GP care is shifted to the profession: an increase in the number of consultations results in lower fees (Vogelaar 2005).

Some experience with risk sharing for follow-up costs is acquired through the early, so-called 'Zaanland system' and 'Amsterdam system' ('Zaanlandse stelsel' and 'Amsterdamse stelsel') in the 1930s and through a bonus/malus experiment in the 1980s.

7.2.2 The 'Zaanland system' and the 'Amsterdam system' [1]

7.2.2.1 'Zaanland'

In 1929, the sickness fund of Zaandam in the Netherlands introduced a new system in which GPs were made responsible for cost overruns in the pharmaceutical budget set by the sickness fund. The system was also supported by the NMG, the Dutch Medical Association (Nederlandsche Maatschappij tot bevordering der Geneeskunst). A reduction in the costs of pharmaceutical care would enable a rise in the payments for medical specialists as well as an increase of the benefit package without rising insurance premiums. Probably, GPs agreed with this system as they wanted to prevent that rising costs resulted in lower fees for their own services. Further, it was in their interests to keep insurance

[1] The information on both systems is based on Van Duuren (1993).

premiums low. Higher premiums could have meant less insured patients and, as a result, less income.

Financial and organisational structures

The *risk package* was composed of drugs. In fact, the physician was free to prescribe the drugs he considered necessary, i.e. no formulary was used by the sickness fund. Most of the available drugs were included in the risk package. Some drugs were excluded, however, like those prescribed by medical specialists and those related to high consumption levels per patient. Pharmacists had to keep track of the physicians' prescribing patterns (volume as well as costs) and had to pass the information to the sickness fund.

The *size of the practice population* is unknown. Presumably, part of the population was insured by one of the other sickness funds or by a private insurer, or had no health insurance at all. It is plausible that the proportion of the practice population with a 'Zaanland' policy differed per practice, but that it was less than 50 percent.

The *normative level of care* was considered to be a rather generous amount of money per patient. About the way it was computed is only known that drug prices were taken into account. Presumably it was based on mean costs and equal for each physician. Of the normative amount, some fifteen percent was paid into an account that was used to pay for the costs of the above mentioned high-use patients. The other 85 percent formed the physician's budget for the risk package. In a later stadium a so-called 'List of Sorts and Quantities' was drawn up on which stipulated the drugs that were permitted and in which quantities they had to be prescribed. Aim was to teach the physicians to prescribe less expensively. It was also decided then to analyse the causes of the deficits more closely.

The system was an example of a cost-decreasing, proportional *malus* system with a norm functioning as a threshold. A deficit was settled by means of a deduction from next-year payments.

The physician's *risk* was *limited* by the above mentioned exclusion of drugs prescribed by medical specialists and those related to high consumption levels per patient.

Results

Due to a lack of information, it is difficult to form a notion of the effects of the system. As far as it concerned drug costs, savings ranged from 11 to 20 percent per year. However, it was suspected that these savings were accompanied by an increase in the costs for specialist care. Possibly, cost shifting occurred by referral to medical specialists of those patients with expected high costs. Another way to keep within the budget was to supply patients with drug samples.

In about 10 percent of the cases, a physician faced a deduction from his payments.[2] Again the information is limited, but some physicians faced cost overruns ranging from 11 to 24 percent. The system seemed to be disadvantageous for physicians with a small practice population as they faced more problems with risk spreading.

Unknown is which consequences the altered prescribing behaviour had for the patients. No information is available about the effects of the system on the patients' health

[2] A 'case' is a budget year for one physician.

status. There is some anecdotal evidence, though, that the benefit package was enlarged or that the premiums remained relatively low.

7.2.2.2 'Amsterdam'

Already in 1853 the board of a sickness fund in Amsterdam concluded that drug costs were too high. Pharmacists were asked to inform the sickness fund of each physician's prescribing pattern. The findings were discussed with the physicians then. For a short period this feedback system was successful but soon drug costs started to increase again. Later, around the same time as the sickness fund 'Zaandam', the sickness fund 'Amsterdam' introduced a new system. Again, the goal of the system was to reduce the costs of pharmaceutical care, but now the final aim was to improve dental care as well as to balance the sickness fund's budget. To a large extent it was comparable to the 'Zaanland system': GPs were made responsible for cost overruns in the pharmaceutical budget. About half a decade after the introduction of the system, the regulation of the sickness fund was modified. At that moment the necessity of the system was restated, but then as way to be able to pay for expensive drugs.

The 'Amsterdam system' was used by 10 sickness funds operating in Amsterdam, but of only one of them the structure and the results are described here.

Financial and organisational structures

The GPs had to use a 'List of Sorts and Quantities' and were also given directions concerning prescription of some special drugs, like insulin. The drugs on the list formed the *risk package* whereas the special drugs were excluded.

The *size of the practice population* presumably differed per practice. Although exact figures are not known, it is plausible that the practice population's proportion with a relevant insurance policy was less than 50 percent.

The *normative level of care* was based on mean costs and was the same for each physician. It was computed quarterly, probably per sickness fund.

Like the 'Zaandam system', the 'Amsterdam system' was an example of a cost-decreasing *malus* system with a norm functioning as a threshold. The difference, however, was that a margin of 10 percent was used. If a physician exceeded the norm with more than 10 percent, the sickness fund could cut his payments. It is unclear whether the malus was proportional to the extent the physician in question exceeded the norm (or the 10 percent margin).

The *physician's risk* was limited as some drugs were excluded from the risk package. Further, a physician with less then 100 sickness-fund insured in his practice was exempted from the malus system. For physicians with a small number of sickness fund patients (< 500) the margins were enlarged depending on their referral pattern. For physicians who referred fewer patients to medical specialists than the mean number of referrals, the 10 percent margin was enlarged as follows:

- to 15 percent (in case of 400-500 sickness-fund insured in the practice),
- to 20 percent (300-400 insured),
- to 25 percent (200-300 insured),
- or even to 30 percent (100-200 insured).

For physicians with a small number of sickness fund patients (< 500) who exceeded the mean number of referrals to medical specialists with less than 50 percent, the 10 percent margin was enlarged as follows:
- to 12.5 percent (in case of 400-500 sickness-fund insured in the practice),
- to 15 percent (300-400 insured),
- to 17.5 percent (200-300 insured),
- or to 20 percent (100-200 insured).

For physicians who had less than 500 sickness-fund insured but who exceeded the mean number of referrals to medical specialists with more than 50 percent, the margin remained 10 percent. Thus the smaller the number of sickness-fund patients in their practice and the more conservative their referral patterns, the more physicians were permitted to prescribe drugs.

The sickness fund 'Amsterdam' applied its system in a more flexible way than did the sickness fund 'Zaanland' by assessing and taking into account the causes for a physician's deficit. Moreover, in the first years the educational function of the system was emphasised. As a result, the 'Amsterdam system' was more popular among physicians than was the 'Zaanland system'.

Results

Apparently, the system had an effect on drug costs in an absolute as well as in a relative sense. For instance, the drug costs amounted to 193,800 Dutch guilders over the first three-quarters of 1931, to 162,400 guilders over the same period in 1932, and to 150,000 guilders over that period in 1937. As a percentage of the premium contributions, the spending on drugs decreased too. In the same years these percentages were 18.32, 15.46, and 13.97 respectively. It is unclear, however, to what extent lower drug prices contributed to this decrease, and whether cost shifting occurred by increasing referrals to medical specialists.

In 1935, 151 physicians (about 56 percent) exceeded the norm, of which 66 physicians (about 23 percent) exceeded the margin of 10 percent.

It is not clear whether the system had a beneficial effect on the health status of the insured. One of its aims, for instance, was to improve dentistry but the effect is unknown. Whether an effect was that necessary drugs were withheld, or that cheaper and perhaps less effective drugs were prescribed, is also unknown. Some patients may have experienced a financial effect of the system. Anecdotal information suggests they had to pay for their drugs themselves if they wanted more expensive drugs than their physician was willing to prescribe.

7.2.2.3 Discussion

Both systems were implemented to reduce the costs of pharmaceutical care by making GPs financially responsible for the drugs budget. Although they functioned in the thirties of the previous century, their experience is still of interest here for several reasons. First of all it illustrates that risk sharing is not a new phenomenon. Secondly, it demonstrates that – although the evidence is limited here – financial responsibility of physicians for

drugs may have an effect on drug costs. Finally, and perhaps most interesting, it highlights some flaws of both systems.

As to the risk package, the systems showed the inherent dangers of restricting it to pharmaceutical care. As specialist care may substitute for drugs, physicians may be tempted to refer instead of to prescribe. Although one should take account of the effectiveness of the drugs during that period of time, it is not inconceivable that pharmacotherapy was often a cost-effective alternative to specialist care. In the 'Amsterdam system' undesirable substitution was counteracted by enlarging the margin (and thus limiting the risk) if a physician had a conservative referral pattern (at least for those physicians with a small number of sickness-fund patients). Another problem was that in the first few years GPs were responsible for the follow-up prescriptions involving drugs initially prescribed by a specialist. In a later stadium these prescriptions were excluded from the risk package, so it was restricted to drugs prescribed by the GP.

As to the size of the practice population, the 'Amsterdam system' recognised the problem of risk relating to practice size. Provisions were made to reduce the risk for those physicians that had a small number of sickness-fund patients.

A no-risk margin may be effective to meet fluctuations in need or demand. The figures mentioned in the foregoing show that it considerably reduced the number of physicians who faced a loss. Adjusting the margin to the number of sickness-fund insured in a physician's practice population may be an effective way to reduce the physician's risk. The inherent flaw, however, is that (risk-sharing) contracts with other sickness funds are not taken into account.

One of the main problems of both systems was the crude method of establishing a norm. As it was based on mean costs and not differentiated for patient characteristics, it was obvious that the prescribing figures of a physician were not solely a reflection of the cost-effectiveness of his prescribing behaviour. Furthermore, as the costs for drugs decreased, the normative level decreased too. The effect of such a system is that it becomes increasingly hard to keep within the prescribing budget and that the incentives to skim cream, to skimp on quality, or to shift costs increase.

Effective or not, the systems were abolished during and after the Second World War. At least three points underlay the abolition. First, during that period there was a considerable delay in the calculating of the physicians' results. Secondly, physicians were dissatisfied with the way the system took account of other determinants of the costs of pharmaceutical care. They argued that their influence on the financial results was limited. A third and maybe crucial cause of the abolition was the enacting of the 'Ziekenfondsenbesluit' (the Sickness Funds Decree) by the German occupiers. As a result, sickness funds were no longer financially responsible but were reimbursed in full by a Central Fund. Hence the GPs as well as the funds had no longer an incentive to contain the costs of drugs.

7.2.3 Bonus-malus experiment 'Tilburg'

In 1984, a Dutch sickness fund started an experiment that aimed at investigating whether the use of financial incentives in the remuneration system of GPs could alter their practice style and result in a shift from secondary care to primary care. Later on, a second

Dutch sickness fund joined the experiment (Van Tits and Nuyens 1987; Van Tits 1989). The ultimate goal was not so much to save costs as it was to strengthen primary care. Further, a goal was to propagate practice patterns that were considered as adequate. To this end, eight primary care practices (13 GPs) formed an 'experimental' group (not randomised but on a voluntary basis).

Financial and organisational structures

The experiment focused on a *risk package* containing five parts: referrals to ophthalmology and 'ENT' ('ear, nose, and throat'), referrals to remaining specialties, hospitalisation days (psychiatry excluded), physiotherapy (exercising therapy), and drugs (prescribed by GPs).[3]

The physicians' public patients (about two-thirds of his total practice population) formed the relevant *practice population*.

The five parts mentioned above were considered as separate cost categories with separately settled *normative levels*. Initially, the norms were based on the figures of all GPs contracted by the sickness funds and defined as ranges with lower limits (i.e. the bonus levels) and upper limits (i.e. the malus levels). They were the same for each practice and not adjusted for differences in the age distribution of the practice populations, distance to the hospital, or any other factors that might have explained differences between practices.[4] The bonus levels were set in a way that 25 percent of the primary care practices were below the levels (in terms of the number of referrals or the costs of drugs). The malus levels were set in such manner that 50 percent of the physicians were above the levels. Within the resulting range no bonuses or maluses were calculated. Later on, the bonus-malus levels were adjusted yearly. To this end, the levels for a practice were corrected for developments in the figures of three comparable reference practices.

The *bonus* amounted to 30 percent of the saved direct costs, the *malus* 30 percent of the extra-generated direct costs. At the end of the budget year, all bonuses and maluses per practice were added up. In case of a positive result, the bonus was paid, but in case of a negative result the physicians did not have to pay anything. In this way the physicians' risk was limited (profit side; see subsection 6.2.1). Nevertheless, physicians could still lose their bonus because of negative results (i.e. maluses) in some parts of the risk package. Besides the range instead of one fixed point, this system resembles the proportional bonus/malus system with threshold as displayed in figure 6.3. It is a cost-decreasing system.

No additional *risk-limiting measures* were taken, as in the end the physicians' risk was limited to the profit side.

[3] A distinction is made between referrals to ophthalmology and 'ENT' ('ear, nose, and throat') and referrals to remaining specialties because the compensation of the GP per referral card differed for both groups of specialties. Psychiatry was excluded as the volume of such care depends on the availability of facilities, and as the length of stay within the facility is unpredictable.

[4] The reason for this was that adjustment was thought of as being administratively complex, and technically problematic. Furthermore, it was considered to be unnecessary as the usual payment system was not adjusted for such factors either (Van Tits 1989).

Table 7.1. Financial and organisational structures of 'Tilburg'

1. Risk package:
   - Referrals ophthalmology and 'ENT';
   - Referrals remaining specialties;
   - Hospitalisation days (psychiatry excluded);
   - Physiotherapy (exercising therapy);
   - Drugs (prescribed by GPs).
2. Size of the population:
   - Public patients (about two-thirds of the practice population).
3. Norm:
   - *First*: based on average cost figures of all GPs having a contract with the two sickness funds; range with lower limit (bonus level with 25 percent of practices below it) and upper limit (malus level with 50 percent of practices above it);
   - *Later*: corrected for developments in the figures of three comparable reference practices.
4. Bonus/malus system:
   - Cost-decreasing, proportional bonus/malus system with range as threshold;
   - Bonus: 30 percent of surpluses;
   - Malus: 30 percent of deficits (solely for calculation purposes).
5. Risk reduction:
   - No additional measures.

Results

The results after four years of experimenting are displayed in table 7.2. It can be seen that the decrease in the number of referrals and the number of hospitalisation days was larger in the experimental group than in the reference group (a decrease of 21.6%, 15.4% and 22.8% versus 13.8%, 3.5% and 13.7% respectively). These results are especially noteworthy because the physicians in the experimental group already had lower figures for referrals, hospitalisation days, physiotherapy, and drug costs than their colleagues in the reference group had. The number of referrals to ophthalmology and 'ENT' decreased continuously across the four years and this decrease was considerably larger in the experimental group. In the experimental group, the decrease in other referrals and in the number of hospitalisation days stabilised at a lower level. In the reference group, the other referrals remained more or less constant, but the number of hospitalisation days kept decreasing. Further, the number of physiotherapy treatments and the costs of drugs in the experimental group increased at a slower pace than in the reference group (7.0% and 20.1% versus 18.1% and 29.3% respectively).

There were no figures on the influence of the financial incentives on the quality of care (for instance in terms of the way physicians practised), the health status of the practice populations and patient satisfaction. In fact, the only indication about the quality comes from the physicians themselves. The majority of them thought that the quality of their practising had improved. They were convinced that they:

- performed more actions by themselves (instead of referring the patients);
- adopted a more critical attitude towards referrals and prescriptions;
- were stimulated to check on their own patients instead of letting medical specialists or the hospital check on these patients.

Table 7.2. Results of the 'Tilburg' experiment

type of care	experimental group	reference group	experimental vs. reference group
referrals ophthal-mology and 'ENT'	-21.6%	-13.8%	-7.8%
other referrals	-15.4%	-3.5%	-11.9%
hospitalisation days	-22.8%	-13.7%	-9.1%
physiotherapy	+7.0%	+18.1%	-11.1%
drug costs	+20.1%	+29.3%	-9.2%

'ENT': Ear, nose, and throat
Source: Van Tits (1989).

Discussion

Due to the small numbers and the non-randomised groups, the research findings are not generalisable. Nevertheless, as the chosen bonus system appears to have had an effect on the behaviour of the participating physicians, the results are of interest here. Not only seems the effect to be reflected in the figures; the participating physicians implicitly confirmed it, given their remarks that the quality of their practising had improved. Although the changes in volumes and costs over the four years may well be the result of the financial incentives, such changes are not necessarily in the interests of the patients. Hence, the question remains whether the bonus-malus system encouraged the physicians to act as their patients' agents. It is difficult to infer this from the experiment, as neither figures on the quality of care are available nor on the extent of cream skimming or cost shifting. The overall conclusion of the research was that, during the experiment, indeed a shift took place from secondary care to primary care and that, in general, the participating physicians altered their way of practising. The apparent shift may be favourable to the agency function, given that care provided in the GPs' practices may be less intrusive and more comfortable for the patients. The altered way the physicians practised was an indication that the agency function had improved by the system. Not only provided the GPs more services by themselves, they also adopted a more critical approach to (repetitive) referrals and prescriptions. They referred more deliberately to specific medical specialists, tried not to lose sight of their patients once they were referred, demanded information on their patients from specialists, et cetera (Van Tits 1989).

Additional judgements of the performance of the physicians' agency function will have to be based on the financial and organisational structures of the bonus-malus system. A first consideration is the composition of the risk package. As it was rather comprehensive, the incentive to alter behaviour was also rather strong. Only the utilisation of some health care goods and services – hospitalisation days for psychiatry, for instance – were *not* affecting the settlement of the bonus. Again, it is not clear whether the comprehensive composition improved the GPs' functioning as agents for their patients. The resulting incentive to reduce volume or costs of the several parts of the risk package may

have had a beneficial effect on the agency relationship, but only insofar as it resulted in a reduction of unnecessary care. It may have had a detrimental effect, however, if it has led to a reduction of necessary care, or if it has resulted in cream skimming.

Further, unjustified care shifting is against the patients' interests. However, as cost shifting was made less attractive due to the comprehensiveness of the risk package, care shifting seemed no major issue. The importance of a careful composition of the risk package is also demonstrated by the problems with the bonus for drug costs. As only costs of drugs prescribed by the GPs were taken into account, physicians had a smaller chance of earning a bonus if they continued drug treatments initiated by the medical specialist. Interestingly, in the 'Zaanland system' and in the 'Tilburg' experiment the GPs' risk was reduced by excluding drugs prescribed by specialists, but obviously at the cost of incentives for substitution. Hence the researchers that evaluated the 'Tilburg' experiment suggested including all drugs into the risk package (Van Tits 1988).

The incentive contract covered about two-thirds of the physicians' practice population. Not known is whether this had a (positive or negative) effect on the treatment of the physicians' private patients.

In the initial calculations of the normative range, the age distribution of the practice populations, distances to hospitals, et cetera were not taken into account. The norms were solely based on the average cost figures of all primary care practices having a contract with the sickness funds. Obviously, this may have resulted in unfairness in the settlement of the bonuses, which on its turn could have prompted physicians to skimp on quality or to skim cream. In subsection 6.4.4, some disadvantages were mentioned of using norms based on actual figures. One disadvantage is the possible continuation of large differences in practice patterns. This was partly circumvented here by using aggregated figures for the norm instead of figures per physician or per practice, and by paying a bonus only if the actual figures were below the bonus level (instead of paying a bonus for each guilder saved). Physicians who had a costly style of practising in the previous years could not take advantage of that by earning a bonus easily. First, they had to reduce their volumes and costs beyond the bonus level in order to earn additional payments.

Another disadvantage mentioned in subsection 6.4.4 is the incentive to increase the costs in the year before the start of the risk-sharing contract. Data collection began 21 months before the start of the experiment, but there was no evidence for such an increase. Besides, in case of calculations based on average costs, the incentive to increase costs in the preparatory year is less strong than in case of calculations of norms based on historical figures per physician or per practice.

Not very encouraging to the agency function seems the chosen bonus system. Although the incentive to reduce volume or costs was limited to thirty percent of the savings, the bonus was not limited to a certain amount. The main constraints in such a bonus system are not provided by the system itself, but may result from the physicians' ethics, the proximity of peers, and the leverage of these peers and of patients. During the experiment a continuous decrease was indeed found for some of the parts of the risk package: for referrals concerning ophthalmology and 'ENT'. For the other parts of the risk package an increase or a stabilisation was found.

The overall conclusion here is that the findings indicate an improvement of the physicians' functioning as their patients' agents. The absolute or relative reductions in the volumes or costs of most of the parts of the risk package formed an important indication. But perhaps more important are the accompanying changes in the way the physicians practised. Nevertheless, the incentives emanating from the chosen financial and organisational structures of the system seem not fully compatible with the agency function. Especially modifications in the bonus system – for instance, an inversely proportional system with a normative target instead of a proportional cost-decreasing system – might have been more effective to support the physicians' agency function as the first system as it makes the desired level of achievement explicit and adverse physician behaviour less likely.

7.3 General Practice Fundholding in the United Kingdom

7.3.1 Introduction

Since the coming into force of the 'National Health Service and Community Care Act', in 1991, several far-reaching changes have been carried through within the United Kingdom's National Health Service (UK NHS). Until 1991, the NHS was for the major part a monopolistic integrated system – in terms of Van de Ven et al (1994). It was mainly tax-financed and a large part (hospital and community care) was provided by public organisations. Family-health services were provided by independent suppliers, but without price competition (the bilateral monopolistic contract model of Van de Ven et al.) A minor part was privately paid for according to a reimbursement scheme.

After the 1991 reforms, the UK NHS mainly got the format of a monopsonistic contract model. Within the still mainly tax-financed NHS, an internal market (a 'quasi-market') was created with separated purchasers and suppliers of care.[5] Different purchasers started to represent the demand side. Health Authorities became the primary purchasers and became responsible for hospital care and community care within their district. A second group of purchasers consisted of GPs that opted for the so-called General Practice Fundholding scheme. The Regional Health Authorities may be referred to as the third group of purchasers, which were responsible for regional care.

Up to 1991, the situation in the UK was roughly comparable to that in the Netherlands. GPs received mainly capitation payments covering GP care. Main differences were that the British capitation payments pertained to the total practice populations (instead of

[5] The creation of a quasi-market is an example of how a regulator (a government, for instance) may restructure the health care system in order to provide a third party with the proper incentives. A quasi-market differs from a conventional market in some of three ways (Le Grand 1991, p. 1260). Firstly, on the supply side there is competition for customers between a variety of not-for-profit organisations and, possibly, for-profit organisations. Secondly, instead of stated in terms of money, purchasing power of consumers is in the form of a voucher or an earmarked budget. Thirdly, the majority of the purchasing decisions are not made by consumers but by agents (third parties) representing them.

a certain percentage of it, like in the Dutch system), and that British GPs received some additional fee-for-service payments and allowances for some specific services.

From the start of the NHS to the 1991 reforms, GPs were in a relatively weak position with respect to hospitals and medical specialists. They could not act as agents on behalf of their patients very well. As far as they did act as their patients' agents, this was mainly with regard to their own care. The absence of any instruments to manage other care hindered them in the execution of their role as gatekeepers, but especially as co-ordinators of care. They had virtually no influence on the cost-effectiveness and quality of other providers' care. A financial risk-sharing system specifically designed to enhance the GPs' agency function, was the General Practice Fundholding system enacted in 1991. By means of fundholding GPs were provided with instruments to improve their agency role. Fundholding rigorously changed the allocation of the risk associated with follow-up care. Participating GPs were allocated a budget to purchase a health-care package on behalf of their practice population. GP fundholding was thus a perfect example of a scheme in which parts of the third-party functions were devolved to physicians. By giving the practices a budget and by providing them with the power and the instruments to purchase health care on behalf of their practice population, the insurance function and the agency function were partly transferred to the GPs.

Insofar as they met certain standards, all GPs could apply for the fundholding status voluntarily. Eventually, about fifty percent of the GPs were involved in the system. In 1999, however, the fundholding system was replaced with a system of commissioning groups called Primary Care Groups (Royal College of General Practitioners, RCGP 1998). Originally, the commissioning groups consisted of GPs that were involved in the planning, purchasing and monitoring of health services. Main difference with fundholders was that initially the commissioning groups were not financially responsible for surpluses or deficits in health care budgets. The new Primary Care Groups, comprising about 50 GPs and serving about 100,000 people, had to start at some point of a spectrum ranging from an advisory role in the commissioning of care by health authorities to full budget responsibility (RCGP 1998, Majeed and Malcolm 1999).

7.3.2 Financial and organisational structures

Originally, the *risk package* of fundholders was composed of:
- all practice team staff costs that were directly reimbursed under standard GP contracts;
- all expenses incurred during management of the fund and other costs associated with participation in fundholding, up to a maximum amount;
- all drugs prescribed and dispensed (within an agreed budget and excluding very expensive patients);
- diagnostic investigations of patients or specimens ordered or performed by the GP;
- initial and continuing outpatient services delivered by hospital-based staff;
- costs related to a defined group of surgical inpatient and day-case treatments (covering most elective procedures);
- costs related to direct access services (for instance, physiotherapy, speech therapy and occupational therapy, dietetics and chiropody);

- health visiting and community nursing;
- elements of mental health and learning disabilities (RCGP 1998).

This was the risk package of the 'standard fundholding' scheme. The package consisted of separated cost categories, but the fundholders were allowed to use the money according to their own view. Excluded from this package were (RCGP 1998):

- accident and emergency services;
- hospital and consultant costs associated with medical cases and non-elective surgery;
- hospital and consultant costs associated with medical inpatient cases;
- maternity services;
- certain preventive and screening tests (like screening programs for breast cancer).

From budget year 1996-1997 on, so-called 'community fundholders' and 'total purchasers' were accompanying the 'standard fundholders'. Community fundholders were not responsible for hospital care. Total purchasers, however, were responsible for almost all hospital care (including emergency care, but with the exception of some rare and very expensive treatments), and for a more extensive part of community health services (NHS Executive 1994). For all the variants held that the local Health Authority purchased services that were excluded from the risk package.

Because of the National Health Service, fundholders had a financial responsibility for the care to be provided to the total *size of their practice population*.

Definition of the *normative level* and thus calculating of the budget posed a problem. The initial idea of central government to uniform the way of financing of Regional Health Authorities, District Health Authorities and Fundholders by making use of differentiated budgets (Department of Health 1989) appeared to be unfeasible, at least at short notice. Not clear was which formula had to be used. Moreover, the use of such a formula would have led to considerable redistribution of means among the practices. This, on its turn, would have implied a threat to the continuity of patient care.

In the first two years of budget setting (i.e. budget years 1991-1992 and 1992-1993), the norm for the hospital care included in the risk package was calculated by using historical costs. This involved the aforementioned problems of data availability, and the incentive to manipulate data. Further, hospitals had problems with the calculation of cost-based prices or were, for commercial reasons, not very eager to publish such figures. Moreover, hospitals faced the incentive to raise prices as this could led to an increase in the budgets that fundholders had available for hospital care (Glennerster et al. 1994). From budget year 1993-1994 on, some Health Authorities calculated the budgets by means of a weighted capitation formula using age and sex, mortality rates (as an index of relative need), availability of beds, self-reported illness, or local unemployment levels (RCGP 1998). These normative levels were not so much binding as they were indicative ('capitation benchmarks') and were to be used in budget negotiations (National Audit Office 1994).

The budgets for drugs were calculated in the same way as the 'indicative prescribing amounts' (later termed 'target budgets') according to the indicative prescribing scheme (IPS) for non-fundholding GPs. Although this IPS was not binding – overspends had no financial consequences – an increasing number of non-fundholders were also allowed to

share in the savings they achieved on drug costs. Regarding the budgets for fundholders, it happened that – because of Health Authorities' policy – they were awarded higher prescribing budgets than non-fundholders (Walley et al. 1995). Owing to the availability of prescribing analysis and cost (PACT) data, the calculation of the drugs budgets posed relatively minor problems. Figures from the previous year were adjusted for changes in age and sex distributions of the practice populations, the number of patients in need of exceptionally expensive drugs, prices and supply of drugs and some other factors (Glennerster et al. 1994).

The budgets for community health services were based on historical figures (Audit Commission 1995). Also for staff costs, budgets were based on figures of previous years. Fundholders were free then to manage their human resources.

In principle, the budget system was an example of a proportional cost-decreasing *bonus system* with a threshold (see subsection 6.4.5). The budget per fundholder formed the norm, and the surpluses (underspends) formed the bonus. Profits could be saved up for four years and had to be invested in patient care. Earning a bonus thus did not result in additional income.[6] Allowed were:

- investments in care included in the risk package;
- the purchase of materials and equipment which:
  - could be used for the treatment of patients;
  - made patient care more comfortable;
  - enabled the physician to practise more efficiently and effectively;
- the purchase of materials for patient information and education;
- investments in premises and surgeries;
- the attracting of new staff.

In case of overspends, the excess of expenditures over income were paid for by the Health Authorities. Overspends could be met by the additional contingency funds that were set aside by the Health Authorities to ensure that fundholders' commitments were met and services to patients could be provided uninterruptedly (Atkinson and Holbourn 1994). As the budgets for fundholders and the overspends were subtracted from the budgets available for Health Authorities, such overspends could reduce the Health Authorities' abilities to arrange health care on behalf of the practice populations of non-fundholders. Health Authorities could attempt to settle deficits in the following years, or they could deprive the fundholding practice in question of further fundholdership (National Audit Office 1994). Withdrawing a practice from fundholding was an option in case the overspending was viewed as due to mismanagement of the fund (Atkinson and Holbourn 1994). Also, Health Authorities were increasingly devising risk-management plans with fundholders. As far as Health Authorities could succeed in making fundhold-

[6] There were some methods to increase income through the fundholding system, though. Lerner and Claxton (1994) pointed at some indirect ways in which GPs could profit from their fundholdingship. Firstly, they could increase their income from the basic payment method (capitation) by attracting more patients to their practices through the use of the budget for services that were attractive for patients. Secondly, they could use the budget to hire extra auxiliary personnel to provide more fee-for-service procedures like immunisations, and to start health promotion clinics. Thirdly, GPs were allowed to invest the surpluses in their practice facilities, thereby increasing their capital.

ers (partially) responsible for overspends, and deficits thus resulted in a malus, fundholders ran a speculative risk. Remarkably, the National Association of Fundholding Practices proposed the formation of some sort of bank from which 'overspenders' could borrow money that had to be paid back within a certain period.[7] This would definitely have altered the fundholders' risk from only profits into speculative.

Several *risk-limiting measures* were taken. First of all, the risk package was designed in such a way that very expensive care was excluded – to which extent differed for community fundholders, standard fundholders and total purchasers though. Secondly, as noted, it was policy of some of the Health Authorities to be rather generous in calculating the fundholders' budgets. Obviously, this decreased the risk of overspends. Thirdly, overspends were not always settled in the following budget years, but were often met by Health Authorities. Fourthly, the fundholders' risk was limited, originally to £5000, later on to a maximum of £6000 per patient per year. Exceeding costs were for account of the Health Authorities. In the terminology of subsection 6.4.6.1, this limitation is called an excess of loss per risk.[8] Fifthly, risk pools were formed. The pools consisted of one or more group practices in such a way that the size of the (combined) patient lists of the practices met the list size requirements. Initially, participation was restricted to (groups of) practices with at least 11,000 patients. But in 1990, thus even before the launch of the scheme, this threshold was already lowered to 9,000 for the first-wave (i.e. standard) fundholders (Maynard 1994; Department of Health 1989). For the second-wave fundholders this was reduced to 7,000 patients, and it was further reduced to 5,000 patients from budget year 1996-1997 on. In the first instance, the restriction for community fundholders was 3,000 patients, but this threshold was abolished from budget year 1997-1998 on (RCGP 1998).

Table 7.3. Financial and organisational structures of GP fundholding

1. Risk package:
   - List of elective hospital care, among which diagnostics;
   - Drugs (prescribed by GP fundholder);
   - Staff costs.
2. Size of the practice population:
   - All patients.
3. Norm:
   - Hospital care: historical figures, later adjusted for factors like patient characteristics;
   - Drugs: based on PACT-data, adjusted for changes in, for instance, patient characteristics and prices.
4. Bonus/malus system:
   - Cost-decreasing, proportional bonus system with threshold;
   - Bonus: 100 percent of surpluses;

[7] The National Association of Fundholding Practices (1996), Risk Management by GP Fundholders – a NAFP Discussion Paper, as cited by the Royal College of General Practitioners (RCGP 1998).

[8] This differs from the terminology used in the United Kingdom, were the term 'stop loss' is used for it. In the reinsurance literature, however, the term 'stop loss' refers to a risk-limiting technique in which the sum of the claims is reinsured as far as it exceeds a set percentage of the premiums (see subsection 6.4.6.1).

   - Malus: depending on what was agreed upon.
5. Risk reduction:
   - List size requirements;
   - Excess of loss per risk of £6,000 per year;
   - Health Authorities responsible for overspends, unless otherwise agreed upon;
   - Non-elective care excluded.

7.3.3 Overall financial results

At the start of the fundholding scheme, several authors pointed out that the size of the fundholding practices would be too small to meet large fluctuations in the nature and the amount of care (see, for instance, Weiner and Ferriss 1990, Drummond et al. 1990, Crump et al. 1991). Nevertheless, during the years the required list size was lowered from 11,000 to 5,000 patients for standard fundholders, and the initial restriction of 3,000 patients for community fundholders was abolished. As a result, the average list size of the participating practices decreased year after year: 12,100 patients in 1991 and 8,400 patients in 1994 (Audit Commission 1995). The financial results, however, turned out less dramatic than some expected. Presumably, this was caused by:

- the fact that fundholders were financially responsible for only a part of total patient care, as expensive and rare diseases as well as accident and emergency services were excluded from the risk package;
- the excess of loss per risk that limited the fundholders' responsibility to the first £6000 per patient per budget year;
- the ample budgets for fundholders, especially for the waves in the first years.

In budget year 1991-1992 the budgets of fundholders in England amounted to a sum of £400 million. In budget year 1994-1995 this had increased to £2,800 million. In budget year 1993-1994 (with a sum of £1,800 million), English fundholders saved £64 million, which was about 3,5 percent. Three-quarters of the budgetholders saved some money, and twenty percent saved at least £100,000. About three percent had a deficit of at least £100,000 (Audit Commission 1995). These surpluses or deficits may, for instance, have resulted from:

- the practice patterns of the fundholding physicians;
- random fluctuations;
- the use of false or incomplete data while calculating the budgets;
- problems with the obtaining of hospital claims.

Generally, Health Authorities attempted to reclaim the so-called 'windfall profits' gained unexpectedly ('by luck') from inaccurate budget calculations (National Audit Office 1994).

Fundholders saved a total of £110 million in the first three budget years. At the end of budget year 1993-1994, 17 percent of these savings were reinvested (Audit Commission 1995). Savings were mainly used for premises (35 percent of the spendings), furniture and equipment (25 percent), medical equipment (18 percent), and the reduction of waiting lists (9 percent).

Glennerster et al. (1994) found substantial differences in the distribution of financial means among fundholders. For practices in three regions, for instance, the budgets for

hospital care ranged from £24 to £106 per patient in budget year 1993-1994. Total budgets ranged from £65 to £203 per patient. It was unclear how far the variations could be explained by the differences in need between the practice populations.

Furthermore, differences were found in the distribution of budgets for hospital care among fundholders and District Health Authorities (i.e. the representatives of the non-fundholders). In one region, for instance, the cost ratios of non-fundholders to fundholders ranged from 0.59 to 0.87 for inpatient care and day-case treatments and from 0.36 to 1.06 for outpatient care (Dixon et al. 1994). Seemingly, these variations could not be explained by differences in need. There were, however, some authors who were critical of the way the costs were calculated in this research (Bowie and Spurgeon 1994, Spenceley et al. 1994). The figures for fundholders and districts were difficult to compare, as the amount of money for non-emergency care available per district depended upon the amount needed for accident and emergency services. This was not the case for fundholders. Further, it was argued that the prices that fundholders had to pay for hospital care were higher than the prices hospitals charged to districts. Such higher prices may reflect the higher transaction costs hospital faced in contracting several fundholders, but they may also be the result of the above mentioned incentive for hospitals to raise their prices in order to cause an increase in the fundholders' hospital budgets. Another reason may be that, because of their size, districts had more market power and, as a result, were more able to negotiate lower prices. However, Propper et al. (1998), who considered District Health Authorities to be captive to their suppliers of care, found evidence contradicting this explanation. To this will be returned at the end of the next subsection.

In spite of the fact that the financial results turned out to be less dramatic than expected by some in the early years, also in the final years of fundholding concerns about the poor budget setting were expressed. Martin et al. (1997) argued that the random variability to which fundholders are subject may lead to inequity within practices (different financial pressures depending on the time of the budget year may lead to different patient care during the year) as well as between practices (different budgetary pressures between practices may lead to different patient care depending on the practice patients belong to). They suggested several strategies:

- Increasing the size of patient groups for which the budgets are set (pooling).
- Setting budgets for a period of more than one year.
- Excluding certain expensive procedures from the budgeting scheme.
- Excluding predictably expensive patients from the budgeting scheme.
- Experimenting with contractual form (i.e. the contracts fundholders conclude with other providers of care).
- Establishing contingency reserves by the Health Authorities to accommodate overspends.
- Careful exploring of variations from budgets.

7.3.4 Results for hospital care

At the end of the eighties, main criticisms of the British health care system pertained to the long waiting lists for elective surgery, to the quality of pathological tests, and to out-

patient care (Glennerster et al. 1994). An important goal of the fundholding system, therefore, was to improve the efficiency of the health care provided by GPs and medical specialists (i.e. the technical efficiency), and to further a more efficient distribution of means among the different parts of the health care system (i.e. the allocative efficiency). Because of the freedom budgetholders had in devising and concluding contracts with hospitals, they were able to make higher demands on the way care was provided in the hospital than they were previously. They also had the possibility to switch between hospitals, or to switch to the private sector. As a result of their increased bargaining power, many budgetholders succeeded in reducing the waiting times for several types of specialist care (Bain 1992, Newton et al. 1993, Corney 1994, Glennerster et al. 1994, National Audit Office 1994).

Referrals

In several studies a decrease in the number of referrals to hospitals was found (Bain 1992, Glennerster et al. 1994, Howie et al. 1994). In another study, no difference was found in the number of referrals for outpatient care made by fundholders and non-fundholders (Coulter and Bradlow 1993). Gogarty and Halliday (1993), however, pointed at the slower increase in those referrals for fundholders. Also in a longer run the number of referrals made by fundholders turned out to have increased less than the increase in the number of referrals made by non-fundholders. In a good three years the increase was 8.1 per 1,000 (7.5 percent) for fundholders and 25.3 per 1,000 (26.6 percent) for non-fundholders (Surender et al. 1995). For two reasons these findings should be interpreted with caution, however. Firstly, only ten fundholders in one region participated in the research. Secondly, out of the six non-fundholding practices, in the end four became fundholder. For three of these four practices, a strong increase in the number of referrals was found in their preparatory year, i.e. the year before they started as fundholder.

As not all of the hospital services were chargeable to the fundholders, it was suggested that fundholding GPs were tempted to have their patients receiving care that remained to be paid for by the Health Authority. It would especially be the case with emergency admissions, for which fundholders were not financially responsible (Weiner and Ferriss 1990). This phenomenon is called cost shifting (see section 6.3). To test the hypothesis that the introduction of fundholding was associated with changing emergency hospital admission proportions, Toth et al. (1997) compared hospital admissions of 21 first-wave fundholding practices (131 GPs) with those of 521 non-fundholders. To this end, they selected four groups of fund procedures for which an emergency admission might be substituted for an elective one, or for which outpatient referrals might be delayed until emergency care becomes necessary. No evidence was found that fundholders increased the proportion of emergency admissions as a means of cost shifting.

In a research among 101 GP fundholders, Whynes and Reed (1994) found that fundholders' intentions with regard to referral behaviour were only to a limited extent determined by the costs of treatments. Fundholding GPs' decisions on referrals for elective surgery in the first place seemed to be determined by their confidence in the ability of the medical specialists, by the length of the waiting times for a first appointment, and by the

level of information from specialist to GP. Relatively little weight was put on the costs of treatments and the style of management in the hospital.

Day case surgery

Fundholders faced an incentive to encourage elective surgical procedures to be carried out as day cases. Day case surgery is less intrusive for patients, may be equally beneficial in terms of outcomes and is, in general, less costly than inpatient care. In a report of the Audit Commission (1996) it was already noted that patients of fundholders were as likely to be treated on a day basis as were patients of non-fundholding practices. Raftery and Stevens (1998) carried out comparative cross-sectional analyses of Hospital Episode Statistics for NHS hospitals in England for budget years 1990-1991 to 1994-1995. Aim was to analyse the influence of target setting and the introduction of GP fundholding on the proportion of elective surgery carried out on a day basis. For a number of procedures, the Audit Commission set targets for the proportion of cases that should be treated as day cases. This reflected the policy objective within the NHS to substitute day case surgery for inpatient treatments. During the five budget years mentioned, the total number of elective surgical procedures increased from 2.7 million to 3.9 million (44 percent). The increase was largely caused by the increase of 117 percent in the numbers of day cases. Four percent increase was found for inpatient based surgery. The proportion of elective day cases increased from 35 percent to 53 percent. In case of fundholders, the total number of elective fundholding procedures increased from 1.9 million to 2.6 million (40 percent). Here too, the increase was largely caused by the increase (109 percent) in the numbers of day cases. Only 0.1 percent increase was found for inpatient based surgery. The proportion of elective day case surgery increased from 36 percent to 54 percent. Although there was a clear incentive for fundholders to increase the relative amount of day case surgery, just a minor – though in most of the years statistically significant – difference was found in the proportion of day cases purchased by district health authorities and those purchased by fundholders. Raftery and Stevens mentioned three possible reasons for the absence of a large difference. A first reason may be that both types of purchasers were equally affected by policies to stimulate day case surgery, like the target setting. A second reason could be that both parties influenced each other's strategies. A third – according to the authors, more plausible reason – may be the emphasis of central policy on day case surgery. This may have resulted in changing behaviour of providers, what on its turn affected both types of purchasers more or less equally. The authors argued that an increase in day case surgery often requires investments in new facilities, like operating theatres. After such large investments, the providers will probably change their behaviour anyhow, irrespective of whether the purchaser is fundholder or not. As the increase in day case surgery was not accompanied by a decrease in the number of surgical procedures carried out on an inpatient basis, their final conclusion was that day case treatments are additional to, instead of a substitute for, inpatient treatments. An alternative to their conclusion may be that the increase in day case surgery was a substitute for an *increase* in inpatient treatments. Nevertheless, the conclusion is shared that it can be doubted that the increase in day case surgery was for a large part the result of fundholders' efforts.

Outreach clinics

Fundholders seemed to be more successful in achieving a rise in the number of 'outreach clinics'. In an outreach clinic, medical specialists perform several diagnostic and therapeutic services outside the hospital that normally are performed within a hospital (Bailey et al. 1994). A part of these clinics were established within primary care practices. Many fundholders succeeded in arranging such clinics (Bain 1992, Bailey et al. 1994, Corney 1994, Glennerster et al. 1994, National Audit Office 1994, Audit Commission 1996). However, the cost-effectiveness of outreach clinics is questionable. The results of a pilot study by Gosden et al. (1997) suggest that additional research into this topic is required. In a (simple) comparative study, they analysed the figures of three dermatology specialists and three orthopaedic specialists, each holding an outreach clinic as well as a hospital outpatient clinic. Only dermatology outreach patients experienced shorter waiting times for their first appointments. Differences were found in travel costs and total costs to patients between the two types of clinics in both specialties, but these were not significant. Due to differences in casemix – in general, outreach clinic patients have less severe conditions – treatment costs were not comparable. Further, staff costs and staff travel costs were significantly higher for outreach clinics than for outpatient clinics, as were the associated opportunity costs – more patients could have been treated if outreach patients were treated on an outpatient basis. The results of a patient satisfaction questionnaire used in the study indicated that patients found the interpersonal nature of the consultation itself more important than access or convenience. Several questions have yet to be answered to be able to assess whether the increase in the number of outreach clinics achieved by fundholders was a blessing or not. Examples of these questions are:

- What is the effect of outreach clinics on GP-specialist communication?
- Do outreach clinics prevent future care and costs through the reduction in waiting times?
- Do such clinics reduce the number of inappropriate referrals?
- Do patients treated within an outreach clinic often still need a referral for tests and treatments (see Gosden et al. 1997, p. 178)?

Substitute services

Difficulties were caused by the services that could be provided by medical specialists as well as by GPs. Fundholders were of the opinion that for those services they could provide at less cost than medical specialists, they had to be able to pay themselves out of the budget. Initially, however, none of such payments were allowed as that was considered to be unfair towards non-fundholders. Moreover, it would be an infringement on the purchaser-provider split. To circumvent this prohibition, several fundholders started new institutions from which they could purchase particular services. From 1993 on, these arrangements were prohibited too. Instead, a list was drafted which contained seventeen services for which fundholders were allowed to pay themselves. Each individual GP that participated in a fundholding practice was allowed to spend thirty hours per month on the listed services (Glennerster et al. 1992, Glennerster et al. 1994).

Price discounts

Fundholders faced a strong incentive to contract providers on the basis of price. Price discounts could result in surpluses or could be used immediately to treat more patients or to treat them in a different way. On the other hand, the incentive was limited due to the way their budgets were calculated (on historic costs). Moreover, price competition was limited by regulations that imposed a 'not-for-profit' as well as a 'no-reserves' condition on NHS providers. Hospitals were required to break-even each year and were not allowed to charge higher prices than in such a way that they covered all costs including depreciation plus a return on net assets of six percent (Propper et al. 1998). Nevertheless, Propper et al. (1998) found evidence that the regulatory rule that prices for procedures had to equal its average costs was broken – at least for some of the eight fundholding procedures for which they analysed price data. They investigated whether market forces had an influence on prices fundholders had to pay for care provided by NHS hospitals during the period 1991-1995. As costs of procedures are unclear, it is virtually impossible to calculate the exact average costs for a procedure, so it is virtually impossible for a regulator to detect price competition based on violations of the regulatory rule. NHS hospitals had an incentive to reduce prices for fundholders in order to attract additional income; this especially in view of the (expected) growth of the fundholding scheme. The largest purchasers (i.e. the Health Authorities) were considered by Propper et al. (1998, pp. 649-650) to be more or less captive to their suppliers (because of inertia, concern about the potential closure of hospitals as a result of moved services, or a lack of incentives). This provided hospitals with the opportunity to increase prices to Health Authorities and to make up for decreased prices to fundholders. A positive association was found for the market shares of Health Authorities with prices to Health Authorities, and a negative association for the market shares of Health Authorities with prices to fundholders, although the latter was significant for only three of the eight procedures. This was especially true for low-cost procedures.[9]

7.3.5 Results for prescribing drugs

In several studies the prescribing behaviour of fundholders was analysed (see, for instance, Burr et al. 1992, Bradlow and Coulter 1993, Maxwell et al. 1993, Dowell et al. 1995, Stewart-Brown et al. 1995, Whynes et al. 1995, Wilson et al. 1995, Harris and Scrivener 1996, Whynes et al. 1997). A rather general finding was that the costs per patient or per 'prescribing unit' were lower for fundholders, or increased relatively slower, than for non-fundholders (Burr et al. 1992, Bradlow and Coulter 1993, Maxwell et al. 1993, Whynes et al. 1995, Wilson et al. 1995, Harris and Scrivener 1996, Whynes et al. 1997).[10] These figures were confirmed by aggregated figures. In budget year 1993-1994, the average drugs costs were £61 for fundholders' patients and £67 for the patients of

[9] Obviously, these findings of Propper et al. (1998) conflict with the remarks some authors made with regard to the article of Dixon et al. (1994). See subsection 7.3.3.

[10] In case prescribing units are used, patients younger than 65 are counted once, whereas patients who are 65 or older are counted three times.

non-fundholders. From the start of the fundholding scheme in 1991, the growth in drugs costs for fundholders seemed constantly lower than for non-fundholders. In budget year 1991-1992, the figures were 12 percent and 15 percent respectively. In year 1993-1994, these were 8 percent and 11 percent respectively (National Audit Office 1994). That fundholders had lower drugs costs had two reasons. Firstly, the lower costs resulted from a decrease or a relatively lower increase in the volume of drugs prescribed (Whynes et al. 1995, Wilson et al. 1995). Secondly, fundholders increasingly prescribed generic drugs (Burr et al. 1992, Bradlow and Coulter 1993, Dowell et al. 1995, Wilson et al. 1995). These figures were confirmed by figures on an aggregated level. In budget year 1992-1993, 52 percent of all prescriptions by fundholders pertained to generic drugs, whereas this was 46 percent for non-fundholders (National Audit Office 1994).

In another study only small differences were found over a period of six months within budget year 1993-1994. For non-dispensing fundholders, the costs per prescribing unit amounted to £21.04, for dispensing fundholders to £20.48, and for non-fundholders to £20.57. Since 1990-1991, the increase in costs was 38.1 percent, 32.0 percent, and 38.7 percent respectively (Stewart-Brown et al. 1995).

A methodological problem for most of these studies is their short period of time. In a study of longer duration, significantly lower average costs per patient were found for fundholders. In 1991 fundholders had on average £2 lower costs per patient than non-fundholders, but in 1994 fundholders spent about £63 and non-fundholders about £81. Remarkably, the lower costs for fundholders were solely the result of lower volumes and not of a profound use of generic drugs (Whynes et al. 1995). Other long-term studies of prescribing costs were performed by Harris and Scrivener (1996) and Whynes et al. (1997). Harris and Scrivener compared the performance of the first five waves of fundholders (totalling 2,649) in England with that of continuing non-fundholders during a six-year period from April 1990. Research questions were whether the prescribing budgets for fundholders led to a reduction in prescribing costs, and whether such an association lasted in the following years. Except first-wave and fifth-wave fundholders, in the preparatory year all groups had slightly higher costs per prescribing unit than non-fundholders, but at the start of their own wave all fundholders had lower costs per prescribing unit than non-fundholders. Over the six years, the prescribing costs of fundholders rose – depending on the wave – by 56-59 percent, while those of non-fundholders rose by 66 percent. The patterns appeared to be similar for all the waves: a small relative reduction in costs in the preparatory year, the largest relative reduction in the first year of fundholding, and a declining relative reduction in the second and third year. After their third year, fundholders faced the same increases in costs, as did non-fundholders. The savings remained, though. Further, as the number of items prescribed per prescribing unit remained stable for each of the waves relative to that of the non-fundholders, the relative cost reductions seemed to be the result of lower average costs per item. These lower averages may have resulted from a more profound use of generic drugs but also, for instance, from lower doses or reduced duration of the prescriptions (Harris and Scrivener 1996). It is unclear what the effect was of the relative reductions in costs and of the lower average costs per item on the clinical outcomes. Whynes et al. (1997) analysed the prescribing costs of – depending on the year – 668 to 696 practices in one region over a five-

year period (budget years 1991-1992 to 1995-1996). Their findings were consistent with those of Harris and Scrivener in the sense that the prescribing figures suggested a cost-reducing effect in the first period and the same increase in costs after that. The timing differed though, as here the effect occurred in the first year of the practices' fundholdership but did not persist beyond that first year. Whynes et al. found no evidence for artificially increased spending by the practices in their preparatory year in order to obtain larger budgets for drug costs.

The idea of the fundholding scheme was that it would stimulate fundholding GPs to prescribe more considerately, thereby putting an end to the rapidly increasing costs of drugs. The influence of fundholders on the overall drugs costs was limited, however. Although they were financially responsible for drugs prescribed by themselves, they had a limited influence on the drugs their patients used. The reduced length-of-stay in hospitals in combination with the fact that medical specialists – forced by hospital budgeting – provided their patients with drugs for only a short period, resulted in a shift of drug costs to fundholders (Crump et al. 1995).

7.3.6 Results for the quality of care

Little is known about the impact of fundholding on the quality of patient care, especially in terms of clinical outcomes (Glennerster 1998). Some aspects of the quality of care, like waiting times, were already mentioned in the foregoing. In one study, the length of consultations in budgetholding practices was analysed. This length was considered to be a measure for the quality of patient care. The average duration of the consultations for patients with pain in their joints remained constant over the period 1990-1992 (Howie et al. 1994). In general, the same held for patients with one of the other of twelve health problems selected for this study (Howie et al. 1995). The patients' satisfaction with the care provided receded a little, though. It is hard to say, however, whether this resulted from the fundholdership of their GPs as the study lacked a reference group.

Another aspect of the quality of care is the accessibility, especially for specific risk groups. Access to care can be endangered by cream skimming. However, this did not seem to be a problem within the fundholding system. The design of the scheme (the risk package restricted to non-emergency and elective care, the norm largely based on historical figures, the excess of loss per risk of £6000 per patient per budget year, and the list size requirements) as well as the fundholders' medical ethics and their reputation may all have contributed to that. This may have been different for the 'total purchasers', but for those fundholders participating in the total-purchasing projects the risk (and thus the incentive to skim cream) was limited by the pooling of large numbers of patients.

7.3.7 Discussion

The lack of proper incentives for parties in the old-style National Health Service to provide health care in a technically and allocatively efficient way and to be responsive to patients, prompted the government to reform the system at the beginning of the 1990s. The question here is whether these reforms indeed improved the performance of the

agency function of those GPs that opted to become fundholder. Overseeing the research evidence, the following conclusions seem justified. As there is no decisive evidence it is difficult to judge the fundholders' functioning as agents for their patients. Moreover, the introduction of the fundholding scheme was just one out of a set of system reforms. The organisation and functioning of Health Authorities and the functioning of other providers of care changed as well. Any evidence should, therefore, be interpreted with caution.

Fundholders appeared to be successful in reducing costs in some of the areas they were responsible for. This seems especially to be true for prescribing costs. Due to the absence of reliable information on the cost-effectiveness of changed (prescribing) behaviour in terms of clinical outcomes, it is hard to say whether cost reductions had a beneficial effect on the patients. On the other hand, it seems plausible that, because of the restrictions on the use of savings, at least part of these savings were indeed used to improve the existing health services, to develop new services, to treat more patients, or to treat them quicker.

Although the fundholding system could have provoked undesirable behaviour, like cream skimming, cost shifting, or quality skimping, no evidence was found of that. There may be a relation between the absence of dramatic changes in the provision of health care, which might have been expected to be a result of fundholding, and the absence of adverse physician behaviour. Stronger incentives imposed on fundholders might have resulted in larger efficiency gains, but also in adverse physician behaviour.

Looking at the structure of the system, the *risk package* can be considered as comprehensive. This is especially true for standard fundholders and even more for total purchasers. A main advantage of a comprehensive risk package is that segmentation of health care is prevented and substitution is furthered. The other side of the coin is that – although depending on the type of care – the more comprehensive is the risk package, the larger is the physician's risk. Adverse behaviour may be stimulated then.

The fundholding system covered each fundholder's total *practice population* and hence the incentives emanating from the fundholding scheme related to the whole of this population. Hence the fundholder can not compensate a loss on the risk contract by increasing the revenues from patients outside that contract.

Clearly, the calculation of a *norm* posed a problem. This may have resulted in an unfair distribution of money, not only among fundholders, but also among the group of fundholders and the group of non-fundholders (i.e. the District Health Authorities). Secondly, the question is whether the requirement of incentive compatibility (see subsection 3.3.3) was met. No evidence was found of widespread cream skimming, cost shifting or quality skimping, but at the same time it is hard to ascertain the extent to which the way of budget setting resulted in an optimal quality and cost-effectiveness of patient care. As to latter, especially the calculation on the basis of historic costs raises doubt.

In general, a cost-decreasing proportional *bonus system* with a norm as threshold is not very conducive to the functioning of fundholding GP's as agent for their patients. The criterion is 'the costs of care' instead of 'the optimal level of care'. Obviously, it depends on the goals one wants to accomplish whether a cost-decreasing system is preferable. The goals of the 1991 reforms were not just limited to cost containment. It is clear that regarding the drugs budgets, the goal was indeed to curb the rising costs. Then, a cost-

decreasing system may, at least for some time, have a beneficial effect. Another important goal was to enable GPs to exert more influence on the quality of non-emergency treatments and outpatient care (including the reduction of waiting lists and waiting times). It is questionable whether the incentive to reduce costs, which emanates from the cost-decreasing proportional bonus system, is the most adequate incentive to achieve this goal. An argument for using this particular incentive scheme is that it stimulates an efficient use of the capacity of health care by which means more patients can be helped. An argument against such a scheme, given the goal just mentioned, is that it stimulates an efficient use of health care a patient receives, but that it does not stimulate to help more patients in the first place. A bonus system with a target might have been more appropriate then.

The set of measures to *limit the physician's risk* seems to have been successful, in view of the financial results. But although the threshold of £6000 (i.e. the excess of loss per risk) may have prevented deficits in the fundholders' budgets, it is unknown how often it occurred that a fundholder missed a bonus because of a few high-cost patients.

It was argued that fundholding resulted in a two-tier system (see, for instance, Bain 1994). This may have found expression in the above-mentioned quicker treatment of fundholders' patients. Perhaps this resulted in longer waiting times for patients of non-fundholders. The findings of Propper et al. (1998), suggesting that Health Authorities had to pay higher prices in order to compensate for the price discounts for fundholders, may be another expression of a two-tier system. Thirdly, it was suggested that fundholders were awarded more generous budgets. The potential threat of a two-tier system would be less of a problem if patients could freely choose between a fundholding and a non-fundholding GP. Given the number and the proportion of fundholding GPs, and given the personal relationship patients usually have with their GP – a characteristic of primary or general medical care – it is questionable whether all patients indeed had a free choice.

Another point of concern resulted from the high administration and transaction costs of the fundholding scheme. Fundholders needed managers and secretarial staff to administer the funds and they had to meet the requirements for computer hardware and software. Health Authorities reimbursed 75 percent of hardware costs, and 100 percent of software costs. Between budget years 1990-1991 and 1994-1995, budgetholders received £165 million for management costs, equalling 2.5 percent of the total amount that was paid to fundholders (Audit Commission 1995). In 1993-1994, 3.5 percent of the total budgets was spent on management, computer systems, and preparing future fundholders (Petchey 1995). This money was withdrawn from direct patient care. These high costs formed one of the reasons for the Labour Government to end the fundholding system in the by then existing form (Glennerster 1998).

7.4 Managed-care experiences in the United States

7.4.1 Introduction

Over the past three decades the American health care system has been facing major changes in the way it has been structured and financed. Noteworthy is the change of the health insurance market. From a system predominated by remote third-party payers, there has been a shift towards so-called 'managed-care organisations', 'managed-care plans', 'alternative delivery systems', or 'integrated delivery systems'. In terms of Hurst, there has been a shift from the reimbursement model towards contract and integrated models.

Managed-care organisations or integrated delivery systems are generic terms describing a myriad of organisation forms. These organisations have in common some degree of integration between the provision and the financing of health care, thereby abandoning the conventional indemnity-insurance way of working.

The first group of managed-care organisations (MCOs) is labelled 'Health Maintenance Organisations' (HMOs). An HMO is an organisation that has a contractual responsibility to arrange and provide a wide range of health care services for a defined population of voluntary enrolled subscribers. These subscribers pay a flat-rated premium. Further, the HMO assumes at least a part of the financial risk associated with health care use (see Luft 1981, p. 2). The prepaid group practice is the oldest and probably most well-known type of HMO. Often, prepaid group practices are subdivided into staff-model, group-model and network-model HMOs. In a *staff-model HMO* physicians are employed by the organisation and only serve the HMO's members. Staff-model HMOs have a closed panel: physicians are not allowed to join the HMO freely, not even if they have patients who are willing to become members of the HMO in question.

In a *group-model HMO* the physicians are employed by the physician-group practice, which on its turn has a contract with the HMO. The group often is a multi-specialty practice. The HMO may pay the group, for instance on a capitation basis or on a cost basis, and the group on its turn pays the physicians. A group model HMO is thus a three-tiered system. As with the staff model, group models are closed-panel HMOs. Physicians in the group may or may not see patients from outside the HMO in question.

A *network-model HMO* is an HMO having contracted several group practices. It is a three-tiered model and may have an open panel or a closed panel.

Besides prepaid-group practices, other types of HMOs can be discerned. In an *individual practice association (IPA) model HMO* the HMO contracts with an association of physicians, the IPA. Affiliated physicians are not working in a group practice and are members of the IPA. This is another example of a three-tiered system. Further, IPA-model HMOs are open panels, so each physician willing to accept the contract terms and having patients wanting to become a member of the HMO is free to join. Usually only a small fraction of the practice population of an affiliated physician is member of the HMO.

A *direct-contract model HMO* resembles an IPA-model HMO in that it contracts independent physicians. The difference is the absence of a middle tier, i.e. the IPA.

In addition to HMOs several other forms of managed-care organisations have been developed. One of these is the *preferred-provider organisation* (PPO), also described as preferred-provider arrangement (PPA), that accounts for about half of the enrolment in MCOs (Smith 1997). Definitions of PPOs differ, which may reflect the various manifestations in practice. According to Enthoven (1985) for instance, a PPO is not an organisation but a type of contract between financiers and providers of care, whereas others distinguish between two-tiered and three-tiered arrangements. A PPA, then, is a two-tiered system while the term PPO is used to denote an organisation between financier and providers (Wagner 1993). Maturi and Rachel (1985, p. 23) gave the following definition of a PPA: 'a health care benefit program arrangement designed to control benefit costs by giving members incentives to use health care providers designated as preferred, but that also provides substantial coverage for services from other health care providers.' This definition seems also applicable to PPOs. Remarkably, the definition emphasises cost containment and ignores quality arguments for stimulating members to use the selectively contracted providers.

PPOs are subdivided into different generations. First-generation PPOs focused on selecting well-known providers of care and using utilisation control techniques. Second-generation PPOs have been focusing on enlarging the number of physicians and negotiating price discounts. The new third-generation PPOs are a combination of the first- and second-generation types in that they try to decrease the number of physicians and to impose price discounts and utilisation management on them (Smith et al. 1997).

An *exclusive-provider organisation* resembles the PPO concept except that health care delivered by other providers is not covered.

In a *point-of-service plan* (POS) members have the choice of receiving care from the selected providers within the HMO or from providers outside the HMO. Coverage is limited in the latter case. When using out-of-plan services, members are confronted with cost sharing and more severe techniques used by the HMO in order to influence the provision of care. In this way, members are stimulated to use the HMO option. Some POS plans, so-called *triple-option plans*, also offer – next to an HMO and a PPO option – conventional indemnity insurance. Characteristic of POS plans is that they offer their members a choice at the point of service instead of at the point of enrolment (Weiner and de Lissovoy 1993).

Although in the United States the relative number of primary care physicians (family practitioners, GPs, general paediatricians and non-subspecialising internists) is decreasing (Starfield 1992), within MCOs primary care physicians are often employed to control the use of health-care services by their members. The vast majority of HMOs employ primary care physicians as gatekeepers (Gold et al. 1995a).

Managed-care organisations differ from each other in the way they are organised. The arrangements between third parties and primary care physicians differ accordingly. Determining the nature of the relationships is, for instance, whether both parties are fully integrated in a corporate sense or only in a functional sense. In the first case (in a staff-model HMO, for instance) the primary care physician is employed by the organisation. The relationship is thus one between employer and employee. In case of just integration

in a functional sense – this may, for instance, be the case in a direct contract model HMO – there is a contractual relationship between distinct organisations. The instruments the third party may employ to further the outcome he aims for, differ in case of a labour contract from those in case of an agreement between two legally separate entities. Furthermore, there is often a correlation between employment and seeing patients from outside the MCO. In staff-model HMOs, physicians only serve patients who are member of the MCO concerned. If the proportion of his practice population that is member of this MCO increases, the primary care physician becomes increasingly dependent upon the third party. This increases the power of the latter and it might be hypothesised that this reduces the need for instruments to be used for influencing the primary care physician. Conversely, a small fraction reduces the primary care physician's dependence and, therefore, makes it easier to switch between third parties. Hence the third party is limited in his choice of contract design and in the use of instruments, as contracts offered by other third parties may be more attractive to the primary care physician.

Also determining the nature of the relationships is whether the managed-care arrangements are made within a two- or a three-tiered system. Within a two-tiered system there is a direct relationship between MCO and primary care physician. In case of a middle tier (in a three-tiered system), the relationship between both parties will only be indirect. The contractual arrangements the MCO makes with the primary care physician in a two-tiered system are now being made with the middle tier. The middle tier on its turn contracts the primary care physician and may thereby change the nature of the arrangements.

Gold et al. (1995a) found that the majority of these organisations shared risks with primary care physicians. In this subsection we discuss two examples. Firstly, the United Healthcare experience in which an IPA-model HMO shared risks with individual primary care physicians. It is a classic example of how a risk-sharing system can fail if it lacks a proper financial and organisational design. Moreover, this experience has been well documented.

Less well documented is the experience of health insurer Blue Cross Blue Shield of Minnesota, which developed a group-model HMO. Only some details are available of its financial arrangements and the organisational structures. Nevertheless, the system is described in this chapter as Kralewski et al. (2000) provided relevant information on financial-risk sharing arrangements between a third party and medical group practices. They also provided information on the payment methods that the group practices used to compensate the physicians operating within those practices. Moreover, the study is unique in the sense that the authors assessed the independent effects of these compensation methods (that is, separate from the effects other managed-care techniques, physician and patient characteristics et cetera may have) on the costs of services for members of Blue Plus.

7.4.2 A primary care network: the United Healthcare experience

In 1974, the SAFECO Insurance Company started an individual practice association of the primary care network type: United Healthcare (Moore et al. 1983, Martin et al. 1985).[11] SAFECO, a large property, casualty and life insurance company, wanted to enter the health-insurance market for small groups. The company was convinced that the physician incentives offered in staff-model Health Maintenance Organisations (HMOs) were the major reason for the lower costs those types of HMOs had. However, SAFECO was mainly functioning outside the large metropolitan areas, which were considered to be the only suitable areas for these HMOs. Moreover, it did not want to provide large amounts of capital to start an HMO. Hence it decided to let United Healthcare share risks with large multi-specialty group practices on a prepaid basis. Having contracted one group practice, it became evident that this strategy had to be abandoned for not enough suitable practices were available. Therefore, the shift was made from a group model network plan to a primary care physician network plan. It started to contract individual primary care physicians then, who had to act as gatekeepers and who had to control the volume and costs of their patients' health care. Every United Healthcare enrolee had to choose a primary care physician and had to visit that physician before going to a specialist or a hospital. To stimulate patients to indeed seek their primary care physician's referral, services were only paid in full with the required approval. As an incentive for the primary care physicians, the risk of their patients' health care costs was shared between them and United Healthcare.

Financial and organisational structures

The *risk package* consisted of all medical care except out-of-area emergencies. It covered services of the primary care physician (i.e. office and hospital visits, office lab and X-rays, and procedures) as well as referred services and hospitalisation (i.e. hospitals, specialists, outside lab and X-rays, and prescriptions). Costs of these services were only deducted from a primary care physician's account if he had approved their provision.

In 1978 only 3.2 percent of the primary care physicians had more than 100 United Healthcare patients in their *practice population*, and about 50 percent had less than 20 of these patients. Although enrolment increased during 1979 and 1980, even then less than 8 percent had more than 100 United Healthcare patients. Unclear is with which proportion of the physicians' practice populations these figures correspond.

The *normative level of care* was based on the average expected costs, adjusted for the enrolee's age and sex. In 1979, for example, the norm for the total risk package ranged from $12 to $40 per month per enrolee.

The United Healthcare plan had a *bonus/malus system* that functioned besides the basic payment system. Basic payments to primary care physicians were on a fee-for-service basis at 95 percent of the charges. Next to this basic system, the physicians shared in the speculative risk of surpluses or deficits in their own accounts at the end of the year. The

[11] For the description of the United Healthcare scheme is drawn on Moore et al. (1983) and Martin et al. (1985).

participating physicians were responsible for fifty percent of the differences between normative and actual costs. This resembles a cost-decreasing proportional bonus/malus system with a threshold, as described in subsection 6.4.5.

Two *risk-limiting measures* were taken. Initially, a physician's maximum bonus or malus amounted to ten percent of his reimbursed charges. From the beginning of 1981, their risk was increased to twenty percent of their charges. Furthermore, a physician was only responsible up to $5,000 per patient per year (excess of loss per risk). Exceeding costs were entirely borne by United Healthcare's reinsurance.

Table 7.4. Financial and organisational structures of United Healthcare' s network

1. Risk package:
   - Primary care services:
     - office and hospital visits;
     - office lab and X-rays;
     - procedures;
   - Referral services and hospitalisation:
     - hospitals;
     - specialists;
     - outside lab and X-rays;
     - prescriptions.
2. Size of the practice population:
   - The majority had less than 100 United Healthcare patients.
3. Norm:
   - Based on the average expected costs, adjusted for age and sex of the enrolee.
4. Bonus/malus system:
   - Cost-decreasing, proportional bonus/malus system with threshold;
   - Bonus: 50 percent of surpluses;
   - Malus: 50 percent of deficits;
5. Risk reduction:
   - *First*: physician's bonus or malus limited to ten percent of his reimbursed charges;
   - *Later*: physician's bonus or malus limited to twenty percent of his reimbursed charges;
   - Excess of loss per risk of $5,000 per year.

Results

The United Healthcare plan was innovative in the sense that primary care physicians acted as gatekeepers and were stimulated to reduce the use and the costs of their patients' health care. Despite this requirement, the plan was successful in attracting enrolees as well as physicians. The benefits were comparable to the comprehensive benefits offered by HMOs. However, as the majority of the primary care physicians were recruited, enrolees did not have to switch to another physician. Primary care physicians, on their turn, were eager to obtain a contract from United Healthcare because they feared losing patients. Another reason was that the minimum reimbursement of 85 percent of the charges (95 percent of the physicians' charges minus a 10 percent loss in case of large deficits) was more than paid by other third parties (Martin et al. 1985, p. 57). Because of this popularity, the United Healthcare plan became the largest of its kind in the United States.

Although successful in attracting enrolees, the large proportion of primary care physicians participating in the network resulted in small proportions of United Healthcare patients in each physician's panel. Hence the impact of the financial incentives was limited. At the end of 1978, seventy percent of the primary care physicians had a surplus in their personal account. The surpluses ranged from $1 to $4,987 with a mean of $430. Only eight percent had a surplus of more than $1,000. Thirty percent of the physicians faced a deficit. These ranged from $1 to $2,877 with a mean of $288. Only three percent had a deficit of more than $500. Moreover, the physicians were only responsible for half of these amounts with a maximum bonus or malus of ten percent of their reimbursed charges.

In the first instance, the overall results appeared to be very satisfactory, but at the end of the seventies use and costs escalated. Over a four-year period (1978 to 1981) total costs per insured increased from $299 to $469 (an increase of about 56 percent). The costs for referral care, drugs, and hospital care increased by 64 percent, 47 percent, and 72 percent respectively. Primary care costs increased relatively slow: by 32 percent (see table 7.5).

Table 7.5. Increase in annual costs per insured

	1978 $	1981 $	Increase %
Primary care	85	113	32
Referral care	89	146	64
Drugs	23	34	47
Hospital	102	176	72
Total	299	469	56

Source: Martin et al. 1985

United Healthcare operated in a competitive market, hence premiums had to be competitive. Compared to national averages, the use and costs at United Healthcare were low. Compared to some of their competitors, however, United Healthcare had a relatively large proportion of patients using outpatient services as well as relatively high costs per user. The use of hospital care did not differ from that faced by competitors, but United Healthcare had relatively high costs per hospital day. In an attempt to control costs, the design of the plan was altered during 1980. The following changes were carried through:

- specialists were contracted selectively;
- primary care physicians were contracted selectively and had to refer to the selected specialists;
- the primary care physicians' financial risk increased from ten percent to twenty percent of their reimbursed charges;

- other managed-care techniques were introduced, like prior authorisation (preadmission review), length-of-stay protocols, outpatient-surgery requirements, and maximum fee schedules for specialists;
- risk-sharing contracts were concluded with pharmacies;
- per diem contracts were concluded with hospitals.

Further, insured faced some coverage changes too, like the introduction of a low-option plan with cost sharing.

In 1981 the costs per enrolee continued to increase, but in 1982 these costs decreased spectacularly. The direct cause of this decrease is unclear. The changes made in the financial and organisational structures may have resulted in a more appropriate use of goods and services. On the other hand, the decrease in costs coincided with a significant decrease in the number of enrolees (and, finally, with the termination of the plan in 1982). Unclear is whether this decrease in the number of enrolees resulted from knowledge of United Healthcare's financial position or SAFECO's decision to sell the plan, from disenchantment with the plan or from a selective disenrollment of high-cost insured as a result of the plan revision. Although there is no evidence to support this, the 1980 reforms may have led to adverse physician behaviour. For instance, the stronger incentives may have prompted the primary care physicians to withhold or to postpone necessary care.

Discussion

The primary reason for SAFECO to enter into health-insurance arrangements was to extend its insurance function by expanding and by diversifying its products and markets. Adding health insurance was seen as a way to increase sales of the primary products (like property insurance) and to meet the demand of clients, and not as a way to develop into an agent on behalf of its insured (in the sense as described in subsection 2.2.3). Consequently, the incentive system was not designed to stimulate the primary care physicians' agency function specifically. It may have had a beneficial effect on this function, though, for instance insofar as physicians reduced intrusive inpatient care by substituting outpatient care or primary care with at least equal outcomes for it.

At most, the goal of the incentive system for physicians was to control costs. SAFECO and United Healthcare intentionally kept these incentives weak, however, as they did not want to deter primary care physicians from joining the plan. They wanted to attract as many physicians as possible for this was viewed as the key to large market penetration. Moreover, SAFECO and United Healthcare were of the opinion that third parties should not interfere in the patient-physician relationship.

As a principal, United Healthcare had three primary strategies to further the achievement of its goals. The first strategy, *selecting physicians*, was only introduced in 1981, some seven years after the start. From 1981, primary care physicians were selected on the basis of the number of (United Healthcare) patients in their practice, their financial results in the previous years, and their willingness to refer solely to the newly selected panel of medical specialists. In 1981 United Healthcare even terminated its relationship with physicians with high costs of care (i.e. 'deselecting' the physicians; see also the description of the managed-care cycle in chapter 5).

For the second strategy, *controlling physicians*, more or less held the same; it was not until 1980 that several techniques were formalised. Besides the already applied techniques of financial incentives and gatekeeping, the most prominent techniques were from then on physician profiling with feedback, prior authorisation, and concurrent and retrospective utilisation review. As noted, the financial incentive for primary care physicians was increased to twenty percent of their reimbursed charges.

Except some analysis of claims, little was invested in the third strategy, *monitoring physicians*. From 1980, the individual accounts of the primary care physicians were monitored. The results were used for selection and feedback purposes. Further, hospitals had to signal the admission of United Healthcare patients.

The design of the major formal technique that was applied right from the start of the scheme, the risk-sharing arrangement, is especially of interest here. The *risk package* was comprehensive and included hospital care and even emergency care within the area. On the one hand, such a design stimulates an integral approach to a patient's health care. It stimulates substitution and it reduces possibilities for cost shifting. On the other hand, it increases incentives for cream skimming and quality skimping. No evidence was found for these forms of physician behaviour, though. In fact, evidence suggested that physicians did not change their practice style once they had entered into a relationship with United Healthcare.

The *proportion* of United Healthcare patients in a physician's total practice population is not known, but the small absolute numbers of these patients – the majority of the physicians had less than 100 United Healthcare patients – in combination with small risks suggest that the incentives to change behaviour were little.

The *normative amount* of money per patient was only adjusted for the insured's age and sex. Such crude norms proved to be insufficient to account equitably for the differences in health status. The combination of practically unadjusted norms, the small proportion of United Healthcare patients, and the comprehensive risk package resulted in account outcomes being mainly determined by bad luck rather than by the physician's behaviour.

It has been noted before that a cost-decreasing, proportional *bonus or bonus/malus system* with a norm as threshold is not very conducive to the physician's functioning as agent for his patients; the criterion is not 'the optimal level of care' but 'the costs of care'. The fact that in the United Healthcare plan the bonus or malus was limited to fifty percent of surplus or deficit hardly improves matters.

The weakness of the incentive system is reinforced by the *risk-limiting measures*. More than the excess of loss per risk of $5,000 per year, it is the limitation of the maximum bonus or malus to ten percent of the reimbursed charges that is responsible. As the total amount of charges was already low due to the small numbers of United Healthcare patients per practice, the resulting risk was not so much an incentive to alter behaviour as it was an acceptable risk of doing business.

Summarizing, the United Healthcare plan was appealing to physicians and enrolees and, as a result, grew rapidly. The major drawback, however, was its inability to constrain the

health-care costs of the insured. The initial strategy of generous benefits, unrestrictive recruiting of physicians and weak incentives turned out to be unsuccessful. The 1980 changes in the plan's structure seemed to sort some effect, but SAFECO did no longer want to carry the huge losses. With hindsight, the plan should have been keener on using several managed-care techniques right from the beginning. Next to stronger consumer incentives, the plan could have benefited from a more selective contracting policy. Further, a combination of several controlling techniques and monitoring could have been used in order to stimulate a more cost-effective provision of care. Moreover, the use of such techniques might have had the effect of self-selection of physicians.

The main conclusions are that:

1. Cost control without intervention by United Healthcare was unsuccessful;
2. Stronger financial incentives were necessary in order to influence the behaviour of the participating primary care physicians;
3. The use of other managed-care techniques besides financial incentives would have made it more likely that the behaviour of physicians was influenced, while at the same time it would have provided a guarantee against adverse physician behaviour as a result of stronger financial incentives.

Noteworthy is the set of strategies that Martin et al. (1985) mentioned as crucial in order to encourage a rational use of health care services by providers (besides another set of strategies to affect the behaviour of enrolees):

- A rigorous cost containment plan that is communicated to providers explicitly.
- Careful selection of primary care physicians and specialists based on quality and costs.
- Provisionary participation in the plan.
- One provider as case manager.
- Individual provider accountability and financial risk sharing.
- Financial incentives that encourage providers to deliver appropriate and cost-effective care.
- Capitation or prospective reimbursement for primary care physicians.
- Utilisation review, especially for hospitalisations and referral care.
- Provider education based on feedback of utilisation review.

In this list the three main categories from the managed-care cycle (i.e. selection, control and monitoring) are clearly recognisable.

7.4.3 Primary care clinics in a Blue Cross managed-care program

Once the SAFECO Insurance Company (see previous subsection) had given up the idea of starting a staff-model Health Maintenance Organisation, it decided to let United Healthcare share risks with large multi-specialty group practices. Due to a lack of suitable practices, however, it ended up with a network plan of primary care physicians. The health insurer Blue Cross Blue Shield of Minnesota, on the other hand, did manage to develop a managed-care program ('Blue Plus') under which they contracted medical group practices. Blue Plus contracted the group practices and paid them in a way that the financial risks were shared between the insurer and the groups. Contrary to the United Healthcare system, the Blue Plus system is an example of a three-tiered system with the

group practice being the middle tier. The enrolees of the managed-care program were required to select a primary care clinic that had to manage their health care. The physicians in the clinics could provide the care themselves, or they could refer the patients to medical specialists.

In studying this program, Kralewski et al. (2000) focused on primary care clinics with at least three full-time physicians. They selected 86 clinic sites with a total of 57,123 Blue Plus members for whom the care was provided and managed by these clinics during 1995. Their selection included primary care clinics that were members of larger group practice systems as well as of independent clinics, and also multi-specialty clinics with a primary care component.

Financial and organisational structures

Risk package

In the study by Kralewski et al. (2000) all the patients' health care costs were attributed to the primary care clinics, even though for a part of the care the patients may have been referred to other clinics. Professional and facility services were included in the study, but mental health care, chemical dependency and pharmacy costs were not.

The contracts medical groups had with Blue Plus during 1995 differed. Kralewski et al. distinguished between contracts with:

1. full-risk capitation for all doctor and hospital services;
2. capitation for doctors' services with some risk sharing for hospital costs;
3. full capitation for all physician services only;
4. full capitation for primary care physician services only;
5. fee-for-service with withhold provision or target rates with settlement at the end of the year or adjustment during the next year;
6. discounted fee-for-service negotiated specifically with the clinic;
7. fee-for-service based on a general fee schedule not specific to the clinic;
8. billed charges (fee-for-service).

Only the first five are risk contracts. Clearly, the risk package differed per contract and ranged from primary care physician services only to all doctor and hospital services. Information on the sources of revenue was obtained from a mailed survey. Apparently, Kralewski et al. were not able to discriminate between the sources of revenue. As a result, the group practice payment data reflected the total practice revenue. Some of the contract forms, therefore, may have been used by other third parties only. Cost data, however, were only obtained from Blue Plus. Kralewski et al. assumed that group practice physicians treat all patients alike regardless of the patients' health plan (and thus regardless of the financial incentives emanating form the several contracts between health plan and group practice). In the subsection 'discussion' is gone into this assumption.

Unclear is whether the physicians' risk packages differed from those of the clinics. This is not likely, however, in view of the small size of the clinics.

Size of the practice population

Firstly, medical groups, including multi-site group practices, were selected if they had a minimum of 200 Blue Plus members who were continuously enrolled during 1995. Secondly, clinic sites with fewer than 50 Blue Plus patients during 1995 were excluded. The rationale for this exclusion was to secure a more accurate assessment of clinic-level costs. As mentioned, the result was that 86 clinic sites with a total of 57,123 Blue Plus members were included in the analyses.

Detailed information on the distribution of the Blue Plus members over the practices is lacking, just like the effect of this distribution on the working and outcome of the incentive system. The lower limit (50 Blue Plus members per clinic site) is known, the number of members per physician, however, is not. Known are the minimum, the average and the maximum number of full-time physicians per clinic: 3, 11 and 79 respectively.

Finally, the proportions of Blue Plus members within the respective total practice populations are not known either. This is unfortunate, for it is not only the absolute number of Blue Plus members that has relevance to the incentive system, but also the ratio of Blue Plus members to the physician's other patients. The incentives emanating from the payment system(s) in use for the other patients or from the contract(s) concluded on behalf of those patients may interfere with the incentives emanating from the Blue Plus contracts (again, see the subsection 'discussion' for the remark about the aforementioned assumption).

Normative level of care

Kralewski et al. did not describe the way the cost levels per clinic were calculated (for instance, based on historical costs). In view of the above mentioned contract forms (see 'risk package') and the fact that the medical group practices had a contract with Blue Plus that placed them at some financial risk, it appears that the norm per clinic was expressed as a cost level.

Within the clinics, norms had to be calculated per (primary care) physician or per group of physicians. Also of these norms, Kralewski et al. gave no description of the way they were calculated. The methods by which the primary care physicians were paid are known, though. From these methods can be deducted that not all norms were calculated as a cost level. Compensation of physicians was based on one or more of the following methods:

1. guaranteed or base salary;
2. individual physician productivity (e.g., cash collection, billings, visits, relative value units, et cetera);
3. individual physician quality of care (e.g., patient satisfaction, chart review, evaluations by supervisor, et cetera);
4. assessment of individual physician management of utilisation (e.g., rate of referrals, laboratory, x-ray utilisation, et cetera);
5. the financial performance of the group of which the individual physician is a member (e.g., share of group net revenue).

Probably, the financial performance of the group was compared with a normative cost level. The normative level is not necessarily a predetermined level, then. It may be a

flexible norm as well and may, for instance, be influenced by the financial performance of other groups. The quality of an individual physician's care, for instance measured by the satisfaction of his patients, will not have been compared with a cost level but with a different kind of norm.

Bonus, malus and withhold

Five out of the eight contract types that Kralewski et al. distinguished at the clinic level, were risk contracts. The practices' financial responsibility can only partially be derived from these contract forms. In case of full capitation, which was the case in contract types 1, 2 (as far as it concerns doctors' services), 3 and 4, the risk is not *shared with* but *shifted to* the medical group practices. Obviously, the practices' risk was not limited, given the term 'full capitation'. In that case, the groups can reduce their risks by increasing the cost-effectiveness of the care, or by cream skimming, cost shifting and quality skimping. In case of contract type 2 (as far as it concerns hospital costs) the risk was actually shared.

In case of the fee-for-service type contract with withhold provision or target rates with settlement at the end of the year or adjustment during the next year (contract type 5), it depends on the exact contract terms whether the risk was shared or just shifted. If, for instance, a practice's risk is the size of the withhold provision, then the risk was shared; costs beyond the withhold are for Blue Plus.

As mentioned, Kralewski et al. were not able to discriminate between the sources of revenue and, therefore, it is not known which contract form(s) Blue Plus used.

At the physicians' level, only the compensation methods are known (see the above). Clearly, ancillary payment methods were used to stimulate the physicians' productivity (compensation method 2), the provision of high-quality care (compensation method 3), or the provision of cost-effective care (compensation methods 4 and 5). Unfortunately, more detailed information on the compensation methods is not available. Not known is, for instance, for which proportion of deficits or surpluses in the risk pool for referral care, laboratory et cetera physicians were responsible, or which proportion of the group net revenue (see compensation method 5) they received. It is assumed here that such variables had an effect on the (costs of) care provided by the physicians in question.

Limitation of the physicians' risk

As mentioned in the above, the risk contracts at the clinic level were mainly full-risk capitation contracts. Only in contract type 2, there was risk sharing for hospital costs. Whether a clinic's risk was limited in another way than by means of a financial responsibility less than a hundred percent is unclear.

At the physicians' level it is known that the risk was spread among a group of physicians (i.e. a risk pool) in compensation method 5, but not *how* the payments were related to the financial performance of the group.

Results

An overall finding was a substantial variance in the mean per member per year (PMPY) costs of care across medical group practices. Even when controlled for patient age, gen-

der, and morbidity (the latter by means of Ambulatory Care Groups), the costs ranged from less than $1,000 to over $3,000 PMPY. The mean is not given but is about $1,550 PMPY (derived from a figure in the article). Ninety percent of the revenue of the more tightly clustered midrange clinics is derived from managed-care programs. Although these programs use financial-risk sharing systems and other managed-care techniques, like profiling or guidelines, there is a $400 difference in PMPY costs within these mid-range clinics.

At the group practice clinic level, the capitation payment methods were collapsed into one variable. Moreover, Kralewski et al. used a corrected capitation payment variable in order to deal with the problem that health plans may favour the use of capitation systems in case of high-cost clinics (a test confirmed this problem). For the correction they used five clinic organisational characteristics that might influence the clinic's ability to deal with the financial risk. These characteristics were:

- group practice or hospital membership (in that case, clinics are expected to be more able to manage risks due to their internal capacity to spread their risk among more patients and because of their ability to control a broader range of providers);
- years of experience with risk-sharing payment contracts (more experienced clinics are expected to be more able to manage risks);
- years of experience of the physicians (more experienced, and presumably more established physicians are expected to favour and to be able to negotiate risk-minimising contracts);
- the number of specialties within the clinic (with more specialists, clinics are expected to be more able to manage risks due to their internal capacity to spread their risk among more patients and because of their ability to control a broader range of providers);
- urban versus rural location (clinics in rural areas, where there is often little competition, are expected to favour and to be able to negotiate risk-minimising contracts).

Using the corrected variable, it was found that capitation had a significant negative effect on costs compared to other payment methods. Interestingly, adding fee-for-service with a withhold provision to the corrected capitation payment variable resulted in a smaller negative effect. Obviously, a fee-for-service system with withhold provision had a different (i.e. smaller) effect on the physicians' behaviour than a capitation system.

At the physician level, compensation based on resource management factors (rate of referrals et cetera) reduced costs significantly compared to compensation linked to some share of the clinic's net revenue. Compared to the latter, both a salary system and compensation based on the individual physician's productivity increased costs. The proportion of a physician's compensation based on his quality of care had a decreasing, but not statistically significant effect on the costs of care.

The effect of the financial incentives both at the clinic level and the physician level should not be judged apart from the context in which they are used. Other managed-care techniques (i.e. selection, controlling and monitoring techniques, see chapter 5) may as

well be used by the third party (in this case the insurer, i.e. Blue Plus, or the group practice). Kralewski et al. found that at the physician level the use of clinical guidelines and physician profiles reduced costs significantly. The use of a computer-based clinical information system or of a more restrictive gatekeeper system had no or no statistically significant effect (see also the discussion section).

Further, some group practice organisational variables were found to have an effect on PMPY costs and may, therefore, be useful as selection criteria. The more experienced the clinic's physicians were, the lower the costs. A higher proportion of women physicians resulted in significantly higher costs, as did, surprisingly, the proportion of primary care physicians.

In the managed-care cycle presented in chapter 5, 'size of the clinic' or 'size of a group practice' is not included explicitly. It is included implicitly, though, as a part of a financial-incentive system in which a clinic or a group practice may function as a risk pool. Kralewski et al. found no significant effect of clinic size on the costs of care.

Discussion

The article by Kralewski et al. is of importance for this chapter; not so much because of a detailed description of a risk-sharing arrangement, but because of the description of the effects of that arrangement. As mentioned, only some details about the Blue Plus arrangement are available. The drawback of the lacking information about the arrangement is that comparison of the Blue Plus system and the other systems described in this chapter is hampered. Moreover, it is difficult to analyse the influence of some specific structures (as defined in section 6.4) of the Blue Plus arrangement on the PMPY costs.

The information in the article suggests that Blue Plus only shifted or shared risks without using other controlling techniques. This points to the fact that although Blue Plus was a managed care program, it primarily functioned as a third-party payer instead of a third-party agent. The capitation payments used by Blue Plus clearly had a cost-decreasing effect, which may be in the interest of insured. This is a crude technique, however, and potentially negative effects of it are not monitored or compensated by other techniques. Not known is whether Blue Plus exerted influence on the contracts that the group practices concluded with their physicians, in order to keep some control over the individual physicians.

On the other hand, the medical group practices displayed features of a third-party agent. They used several techniques like financial incentives, primary care gatekeepers, guidelines and physician profiles. Not all of the financial and organisational techniques had a cost-decreasing effect. Kralewski et al. concluded that the most cost-effective clinics were the smaller multi-specialty clinics that used guidelines, profiles and compensation systems based on resource management factors and in which the physicians shared the group net revenue. As, moreover, these clinics are often owned by the physicians, they have the attributes of small family-owned businesses.[12]

[12] In a family owned business the goals of the organisation and the employees may converge. Further, the assumption of opportunistic behaviour, which is at the heart of the agency theory, may have to be

Whether the arrangements between clinic and physicians were beneficial to the Blue Plus members in another than a cost-effective way is unclear. Assuming that the *risk package* of a clinic and the individual physicians were the same, this package ranged from only primary care physician services to all doctor and hospital services. The larger risk package with all doctor and hospital services will make it more difficult to shift costs.

The *size of the practice population* may pose a problem if it is as small as it was in at least some of the clinics. Unless the other insurers with whom a physician or his clinic had contracts used the same incentives, a small proportion of Blue Plus members may have diluted the incentives emanating from a physician's risk contract. This was also the case with the United Healthcare plan (see the previous subsection). For instance, the fact that Kralewski et al. found no effect from gatekeeper systems might have been the result of the other physician's patients being free to make appointments to see (some) specialists without a referral.

On the way the *normative levels of care* were calculated no information is available, except that other things than costs were also taken into account.

Of the *bonus, malus and withhold* systems is only known that at the clinic level at least in some contracts full capitation was used as payment system. Although a clinic may be able to bear such risks, the fact that the clinics were small implies that even at the clinic level such a system may provoke undesired behaviour. This is especially true if the capitation system is used for expensive care. Moreover, as in some clinics the physicians' compensation was based on the net revenue, such a payment system may even provoke undesirable physician behaviour. Risk sharing reduces such incentives.

Especially in case of full capitation for high-cost care, some measures to *limit the physicians' risk* are desirable. An effective mechanism is the use of risk pools, which is used in some of the contracts between clinic and physicians.

A final remark is about the assumption Kralewski et al. made. They assumed that group practice physicians treat all patients alike regardless of their patients' health plan. The assumption was made to overcome the problem that they were not able to discriminate between the sources of revenue, whereas the cost data only related to patients enrolled in the Blue Plus program. The authors did not go further into the subject, but this assumption seems doubtful. If it is correct though, than this equal treatment may result from the fact that physicians:

- are insensitive to the different incentives emanating from different contracts;
- are only sensitive to the incentives emanating from the contract they concluded with the group practice, but are insensitive to the different incentives emanating from the different contracts the group practice concluded with the several health plans;
- choose an 'average' treatment pattern in a way the effect from the several incentive systems is optimised;
- choose a way of treatment in response to the most attractive contract and then choose this way for all their patients;

---

relaxed in such businesses. See also subsection 3.2.3.

- or may have different contracts of which the incentives have the same effect on their behaviour.

7.5 Conclusion

7.5.1 The five aspects of financial-risk sharing

In this chapter we analysed several arrangements in which the financial risks of follow-up costs are or were shared between third parties and (groups of) GPs. We use the analysis to provide an answer to the first research question in this chapter:
What are actual effects of different systems of financial-risk sharing on the performance of general practitioners?

Risk package
In the examples described here, the risk package differed considerably and ranged from drugs only ('Zaanland' and 'Amsterdam') to an extensive package consisting of primary care services, several referral services and hospitalisation (like GP Fundholding or United Healthcare). An extensive package has the advantage that it stimulates an integral approach to health care, stimulates substitution and reduces the possibility to shift costs. The Dutch systems with the limited risk packages illustrated the difficulties of composing the risk package. On the one hand the Dutch systems showed that the care included in the package should be restricted to care on which the GPs can exert some influence. On the other hand, these systems showed that since only costs of drugs prescribed by the GPs were taken into account, physicians had a smaller chance of earning a bonus if they continued drug treatments initiated by the medical specialist. By excluding the drugs prescribed by specialists, the GPs' risk was reduced. Obviously, this was at the cost of incentives for substitution. Hence the researchers that evaluated the 'Tilburg' experiment suggested including all drugs into the risk package (Van Tits 1988).

GP Fundholders (standard fundholders and especially total purchasers), United Healthcare physicians and Blue Plus physicians were responsible for extensive risk packages. Evidently, in most schemes drug costs were included. Apparently, especially this part of the risk package is considered to provide room for improvements in the cost-effectiveness or quality of care.

Size of the practice population
Regarding the size of the physician's practice population and the proportion for which he is at risk, large differences were found. Primary care physicians participating in the United Healthcare plan, for instance, had only small proportions of relevant members within their practice populations, while GP Fundholders were responsible for all their patients.

The United Healthcare experience showed that due to such small proportions the incentives were limited, as were the effects. Obviously, the proportion of a primary care physician's practice population for which the third party concludes a risk contract has to

be substantial. But it does not have to be a hundred percent, like in the fundholding system, to have an effect though. In case of the 'Tilburg' experiment, for instance, about two-thirds of the physicians' practice populations concerned publicly insured patients for which they had a risk contract. This proved to be enough to have a significant effect on the physician's behaviour.

Normative level of care
Undoubtedly the most problematic aspect of the risk-sharing arrangements proved to be the calculation of a norm. In the absence of a more sophisticated way to establish a normative level of care, historical figures were used or crude adjustments (for example, for age only) were made to average figures. In all systems the norms resulted in uncertainty about the cause of a difference between norm and actual figures (i.e. volumes or costs). For instance, patient characteristics may have played a larger role than the physicians' behaviour (see, for instance the Dutch systems). In case of a norm based on historical figures, physicians were rewarded for inefficient behaviour in the previous years (see, for instance, 'Tilburg' or the GP Fundholding system).

Bonus, malus and withhold
The logical consequence of the problems with the calculation of a norm is that problems arise once bonuses and maluses have to be calculated. In all the schemes cost-decreasing bonus or malus systems were used, mostly in a proportional form. The exact form differed, though. In some schemes a margin had to be exceeded before the bonus or the malus was applied ('Zaanland' and 'Amsterdam').

Further, the percentages that were used to calculate a bonus or a malus based on the difference between norm and actual figures differed too. United Healthcare shared the risks with the participating physicians on a fifty-fifty basis. GP Fundholders, however, were responsible for hundred percent of the difference as were some of the clinics participating in the Blue Plus scheme.

Limitation of the physician's risk
Besides by a bonus/malus system in which the risk for United Healthcare physicians was reduced to fifty percent of the surpluses or deficits, the total financial risk was limited to ten percent, and later to twenty percent of the reimbursed charges. Furthermore, the annual financial risk was also limited to $5,000 per patient (an excess of loss per risk). GP Fundholders faced a different system and had a full responsibility for the first £6,000 per patient per year (excess of loss per risk). The Dutch 'Amsterdam system' had the most refined risk-limiting measures; the no-risk margin used within the malus system was inversely related to the number of insured. In general, however, the risk-limiting measures were rather simple as compared with the systems described in subsection 6.4.6.

7.5.2 The effects of different systems of financial-risk sharing

The majority of the experiences described in this chapter at least indicate an effect of risk-sharing arrangements on the behaviour of GPs or primary care physicians:

- The 'Zaanland system' and the 'Amsterdam system' both showed savings in the drug budgets of the sickness funds.
- In the 'Tilburg' experiment, the experimental group demonstrated a larger decrease in the number of referrals and hospitalisation days than the reference group did. Further, drug cost and the number of physiotherapy treatments increased less.
- GP Fundholding appeared to be successful in cost reduction, especially with regard of drugs, reduction of waiting times for specialist care, decreasing the number of hospital referrals, increasing the number of day case treatments et cetera.
- Blue Plus resulted in cost reduction, both at the plan level and the physician level, although the effect differed per payment method.

United Healthcare was not successful in altering the physicians' behaviour initially, but after structural changes were carried through costs per enrolee decreased spectacularly. It is not clear though whether the modified arrangements induced the physicians to alter their behaviour or whether it, for instance, resulted in a selective disenrollment of high-cost insured.

But despite the evidence, there is no clear answer to the question whether there is a differential effect of the distinct systems of financial-risk sharing on the performance of GPs (as agents for their patients). Firstly, the behaviour of a GP will probably not result from a risk-sharing arrangement only. Financial-risk sharing for follow-up costs is just one financial incentive a physician may face, whereas the use of (financial) incentives is just one possibility out of a set of control techniques that the third party has at its disposal. Furthermore, the physician may respond to the other steps of the managed-care cycle, (re)selection and monitoring, or to stimuli from outside the third party-physician relationship.

Secondly, in the analyses of risk-sharing systems the dependent variable is often 'the costs of care' or 'the volume of care'. The costs or the volume of care, however, give only a partial picture of the effect of a risk-sharing system on the GPs' functioning as their patients' agents. To judge this more closely, other effects are equally important, like the quality of patient care. Hence it is important to evaluate whether undesired physician behaviour, like cream skimming, cost shifting or quality skimping, is provoked by the risk-sharing arrangement.

About managed care Steiner and Robinson (1998, p. 178) concluded that 'despite literally thousands of publications since 1990 whose subject is some component of the managed care approach, it is still not possible to answer fundamental questions about the independent contribution of each component to organisational performance. There are almost no randomised-controlled trials of these techniques in managed care settings. Most publications either describe or advocate the use of techniques, without any evidential basis; many others evaluate interventions only in qualitative terms, lack comparison groups, and make no tests of statistical significance'. The number of publications on financial-risk sharing may be smaller, but the conclusion is comparable. Experimental research may provide some answers, but because of the interaction with other incentives the physician faces and with context variables, a decisive answer to this chapter's first research question is unlikely.

7.5.3 The framework

By describing and analysing the examples using the analytical framework of risk sharing developed in chapter 6, the framework is put to the test here. Thereby, an answer is sought to the second research question in this chapter:

Does the analytical framework of financial-risk sharing sufficiently provides insight into the key differences of systems in which the risk is shared between third party and general practitioner so as to infer the effectiveness of such systems?

The framework proved to be useful in analysing the risk-sharing systems systematically. It helped to gain insight into the key aspects of several systems and provided a basis to compare them. Ideally though, researchers in this field should apply the same framework while analysing a system and while reporting on it. Now the framework had to be used to review sources providing information that was not acquired and reproduced according to the framework in question. In general there was enough information available to derive the characteristics of the systems sought after, though. For the Blue Plus system, however, this proved to be difficult. That it did not work out in the latter case seems not so much the result of the framework as it is of a lack of information.

8 TOWARDS A SYSTEM OF FINANCIAL-RISK SHARING

8.1 Introduction

In chapter 6 we argued that there are good reasons for sharing the financial risk between third-party agents and GPs. Hence the question arises whether some lessons can be derived from the analysis of the theoretical and the practical models of financial-risk sharing. These lessons can not be definitive enough to form a normative model of financial-risk sharing. First of all, only a rather limited number of practical examples are described here. They are primarily meant to be an illustration and do not provide a representative description of the arrangements in the real world. Moreover, the examples are purposely picked from several health care systems. This provides a broad picture of the use of financial-risk sharing, but makes it difficult to derive conclusions that are applicable to other health care systems. Further, a crucial factor in the design of a model of financial-risk sharing is the resulting amount of risk that is transferred from third-party agent to GP. The five main aspects of financial-risk sharing all determine this risk and have to be balanced accurately. Quantitative analysis is crucial then. A final argument is that there will not be one definitive form of financial-risk sharing. What is optimal will differ per country and per health care system, will depend on the position and the role of the GP, will depend on the (political) goals one wants to achieve, will depend on the preferences of the GP, et cetera.

In spite of these limitations, some lessons can be learned indeed by which we will attempt to answer the following research question:

How should systems of financial-risk sharing be structured?

8.2 The five aspects of financial-risk sharing once more

Risk package

The size and the nature of the risk package determine the physician's risk and by that the incentives and possibilities to provide cost-effective and high-quality care. It also determines the incentives and possibilities to skim cream, shift costs or skimp on quality (i.e. to show adverse behaviour).

In chapter 7 we showed that the risk package differed considerably within the several examples, and that it ranged from drugs only to a very broad package consisting of primary care services, referral services and hospital care. Clearly, an extensive package promotes integrated health care and enhances substitution. It also reduces the possibility to shift costs, although the incentives to do so may increase. The larger the risk package,

the larger the physician's risk may become. The larger the physician's risk, the more he may want to reduce his risk, for instance by means of cost shifting. However, a larger risk package makes it more difficult to shift costs.

For some care it will be easier to diagnose, to estimate the costs of treatment and to exert some influence on the way the care is provided than it will be for other care. This is especially important if the GP has to arrange and pay for follow-up care, i.e. is a purchaser of care. Although there may be good reasons to start with a very limited risk package (only drug costs, for instance in order to gain experience), in the end the physician should be responsible for a broad risk package (like in the GP Fundholding system) in order to prevent cost shifting and to facilitate substitution. Very expensive and 'open-ended treatments' should be excluded though. The larger risk because of a more extensive risk package can be reduced by means of one of the other aspects of risk sharing.

Finally, the risk package can be divided into separate cost categories, but that is merely a refinement of a system after some years of experience. Advantages of such a categorisation are an increased insight into the costs and the option to vary the bonus system per type of care. The latter is administratively more complex though. Moreover, if the categorisation and the bonus system are not carefully balanced, the result may be reduced substitution or unwanted substitution of expensive care for inexpensive care (if the physician runs a larger risk for the inexpensive care).

Size of the practice population

It may be hypothesised that the incentives are likely to be too limited if the GP is not at risk for a substantial part of the practice population. What is substantial, however, will also depend on the rest of the incentive contract. If, for instance, the risk for a minor proportion of the practice population is large enough, the physician may still be inclined to behave in the desired direction. The less than 100 United Healthcare patients that those physicians had within their practice population turned out not to be enough – especially in combination with the small risks. The about two-thirds of the practice population within the 'Tilburg' experiment, however, proved to be substantial enough to influence the participating physicians' behaviour. Hence there will be a trade-off for the third party between the proportion of its members within the physician's practice population and the strength of the incentives emanating from the other aspects of the risk-sharing arrangement.

Another aspect that the third party should take into account (if known) is what kind of arrangements other third parties made with the GP. Other contracts can dilute the effect the third party may want to have.

Besides that the proportion has to be substantial in order to draw a reaction from the physician, a larger group of patients makes the physician less vulnerable to random fluctuations in health-care costs. Although depending on the risk package, the number of patients will have to be larger than the size of an individual physician's practice population though. Hence pooling or other forms of risk-reduction will be necessary then.

Finally, a larger proportion of the population for which the physician is at risk makes it more difficult to shift costs to other third parties.

Normative level of care

At the heart of a risk-sharing arrangement is a norm. Historical data may be used or crude adjustments (for example, for age and sex only) can be made to average figures in case a more sophisticated way to establish a normative level is absent. Advantage of the use of historical figures, though, is the relatively smooth transition from a system in which the third party bears all the risk to a risk-sharing system. Disadvantage is that physicians are rewarded for inefficient behaviour in the previous years. Furthermore, a difference between norm and actual figures (i.e. volumes or costs) may very well result from factors exogenous to the physician. Patient characteristics, for instance, may have played a larger role then than the physicians' behaviour.

The norm should at least meet two requirements. The first one is the requirement of fairness. The norm should account for those systematic factors of which the resulting costs can not be influenced by the physician. Secondly, the norm should meet the requirement of incentive compatibility: the arrangement should provide the proper incentives instead of provoking cream skimming, cost shifting or quality skimping. This asks for a more sophisticated norm within risk-sharing systems. Hence econometric modelling is vital. However, as in health care systems with a risk-adjusted capitation formula for calculating the health insurers' budgets, the calculation of a risk-adjusted norm for physicians proved to be difficult. The Fundholding experience has been a good demonstration of the permanent struggle to calculate a risk-adjusted capitation payment for practices participating in the scheme. If the bonus or malus is related to the difference between actual and normative costs, then the difficulties of calculating a norm will have consequences for the calculation of bonuses and maluses.

Of course, there are alternatives to econometric modelling. Norms can be specified as a set quality level, like *x* percent compliance with guidelines or protocols. Norms may also be specified as a certain volume, like the number of referrals or prescriptions per patient per year. The scope of guidelines or protocols is, however, often limited to a set of specific diseases. Setting a norm in terms of a certain volume is an option. National or regional figures may be used, but again corrections have to be made for factors over which the physician has no control. Further, the GP can only influence the volume of care as a way to reduce his financial risk. In case of a norm expressed as a cost level, he may try to influence the volume as well as the price of care, or attempt to substitute relatively inexpensive care for relatively expensive care. There is increased scope for the kind of behaviour of the GP the third party may aim for.

Bonus, malus and withhold

A bonus system may provide a cost-increasing or a cost-decreasing incentive, depending on whether actual costs should be higher or lower than the norm in order to receive a bonus. In addition, a bonus may be proportional to this difference (the larger the difference, the larger the bonus) or inversely proportional (the more equal actual and normative costs are, the larger the bonus). From a third-party agent's point of view an inversely proportional system with a target may be preferable to a cost-decreasing proportional system with a threshold, for in the first system the norm makes the desired level of achievement explicit. The bonus is at its maximum if the norm is reached and the danger of adverse

physician behaviour is reduced to a minimum. A cost-decreasing system makes the desired level of achievement less explicit and stimulates the physician to withhold care completely. Determining the optimal norm, however, is difficult.

For proportional bonus systems holds that, since in risk-sharing systems the risk is shared between third party and physician, the bonus has to be less than hundred percent. Only if risk-limiting measures are taken, then the bonus can equal the difference between norm and actual level. Hence the choice for a certain percentage should at least be related to the choice of the risk-limiting measures.

Another consideration for a third party in designing a bonus system is whether the bonus is varied per type of care. It is a way to refine the system and to adjust the amount of risk, for instance to the extent the GP can influence specific care or to the risk associated with the type of care. Hence the design of the bonus system should particularly be related to the design of the risk package. Disadvantages of a variable bonus system are the inherent danger of reduced or even undesirable substitution, and its complexity with connected administration costs. But if the risk package consists of different types of care and if the risk differs considerably per type of care, a variable bonus can be a useful refinement. Otherwise, a few episodes of an expensive disease may ruin the incentive system.

Limitation of the physician's risk

In general, the risk-limiting measures found in practice are rather simple as compared with the possible coinsurance or reinsurance systems described here. The use of one or more measures is necessary though, especially in view of the fact that the risk package and the bonus have to be substantial in order to reduce the possibility that cost shifting occurs and to stimulate the physician to behave in the desired way.

Coinsurance by means of a risk pool is a good way to increase the size of the population for which the physicians are financially responsible, to spread the risk over a group of physicians and to make the physicians less vulnerable to random fluctuations in health-care costs. In determining the size of the risk pool, risk reduction has to be balanced against financial incentives to behave cost-effectively. To retain these incentives, a physician group with a maximum of about ten physicians may well do. Each physician is responsible for ten percent of the surpluses or deficits then.

The maximum risk can and has to be bounded by upper limits, for a few expensive patients at the beginning of a financial period may ruin the incentive system or may even ruin the participating physicians. To this end, one of the reinsurance techniques has to be incorporated in the risk-sharing arrangement. Given the physicians' limited capacity to bear financial risks, an 'excess of loss per risk' and a 'stop-loss contract' provide the best assurance. Alternatively, a 'quota-share arrangement' or a 'n largest claims' provision can be used.

8.3 Discussion

In answer to the tenth research question:

How should systems of financial-risk sharing be structured?

we recommend to start carefully. Then an initial risk-sharing arrangement can have the following characteristics:
- A limited risk package, for instance, consisting of drug costs only.
- A risk contract for at least fifty percent of the practice population.
- A norm at least partly based on historical figures in order to enable a relatively smooth transition from a risk-free to a risk-sharing system.
- A simple bonus system. Whether it is a cost-decreasing proportional system or an inversely proportional system with a target depends on the third party's goals. In case of a proportional system, the bonus (and malus) should be substantial, but should also be substantially less than hundred percent of the difference between norm and actual costs. If profit or loss is limited to a fixed amount, then a bonus (and malus) of around fifty percent results in an incentive that maintains longer than a bonus (and malus) of hundred percent.
- An 'excess of loss per risk' in combination with a 'stop-loss contract'. A group practice or a locum group, for instance, are an obvious choice to form a risk pool.

If a 'smooth' transition has been made to a risk-sharing system, if third parties as well as physicians have gained experience with risk sharing and if the effects on behaviour and outcomes are known, then a less conservative arrangement can be considered. The arrangement may then have the following characteristics:
- A broad risk package from which expensive and 'open-ended treatments' are excluded.
- A risk contract preferably for the whole practice population.
- A norm that functions as a target and is risk-adjusted or based on an optimal level of care (or a combination of both).
- A more sophisticated bonus system with a variable bonus. A third-party agent will probably prefer an inversely proportional system with a target as it makes the desired level of achievement explicit and adverse physician behaviour is less likely. Again, in a proportional system the bonus should be substantial but substantially less than hundred percent, for instance fifty percent.
- An 'excess of loss per risk' in combination with a 'stop-loss contract'. A group practice or a locum group, for instance, are an obvious choice to form a risk pool.

The five main aspects of a financial-risk sharing arrangement have to be balanced carefully. Each aspect partly determines the risk that is transferred from third-party agent to GP, but it is the ensemble of these aspects that is decisive. Altering the risk package or the population at risk will have consequences for the design of the bonus system and the

risk-limiting measures. Quantitative analysis is a crucial step then in the design of such arrangements.

Besides the five main aspects of financial-risk sharing arrangements, third parties may take some additional aspects into consideration while designing an arrangement. The bonus and the malus may be maximised. Further, the physician may be restricted in the use of the bonus. It may be arranged that the physician has to invest the bonus in patient care, instead of using it as additional income.

Chapters 6, 7 and 8 are devoted to financial-risk sharing arrangements. Financial-risk sharing, however, is not used as an isolated technique. It is one form of control by incentives (i.e. financial incentives), whereas controlling a physician is just one of the three phases of the managed-care cycle. Hence the third-party agent will and has to use financial-risk sharing in combination with other managed-care techniques. Initial selection strategies can be used to select cost-conscious and high-quality providers. Other controlling techniques can be used complementary to financial-risk sharing. Further, the third party has to monitor the physician's behaviour and the outcome to know the effect of the arrangement. The acquired information can then be used as feedback and in re- or deselecting decisions.

9 SUMMARY AND CONCLUSION

9.1 Introduction

Is there a rationale for financial-risk sharing between third-party agents and general practitioners? And if so, how should systems of financial-risk sharing be structured? These are the two central questions of this research. These questions are relevant since incentive systems are viewed as an effective way to increase the efficiency of the health-care sector. As part of ongoing health-care reforms, third parties, like public or private health insurers, are stimulated to organise a more efficient health-care system of higher quality. Often, third parties are made financially responsible for a specified package of health-care goods and services for a defined group of members (i.e. insured) and for a certain period of time. Third parties on their turn can use techniques to achieve the goals of efficiency and high-quality care. Because general practitioners (GPs) have a large influence on the nature, the quantity, the quality and the costs of care, third parties may consider making arrangements with these physicians. Part of the arrangements may be the use of financial incentives, like the application of financial-risk sharing.

The behaviour of physicians partially determines the efficiency of the health care sector. Given the worldwide interest in the efficiency of this sector, it is remarkable how little reliable empirical research has been done into the effect of financial incentives on the behaviour of physicians and the outcome of their actions. Also, little theoretical, conceptual research has been done. This holds especially for payment systems for GPs and for systems of financial-risk sharing. The main purpose of our study is to construct a conceptual framework for systems of financial-risk sharing between third parties and GPs. We will use this to analyse several examples of financial-risk sharing. Then we will see which lessons can be learned for the structuring of systems of financial-risk sharing. By this, we aim at contributing to the literature on payment systems for GPs.

Financial incentives are just one possible technique out of the set of potential techniques the third party may use to try to influence the behaviour of GPs. In health care (the use of) such a set of techniques is often designated as ‘managed care’. Managed care is a rather diffuse concept that lacks a sound theoretical framework. Hence before creating a conceptual framework for systems of financial-risk sharing, we consider it necessary to provide managed care as a whole with a theoretical basis and a clear classification. By this, we aim at contributing to the literature on managed care.

In this research we used agency theory to analyse the relationship between third parties and GPs. There are several reasons why we chose an agency perspective. A first reason is that agency theory seems to correspond well with the situation central in this thesis: the contractual relationships between two parties (that is, third-party agents and GPs) in which the one party (the principal) enters into a relationship with a second party (the agent) in the expectation that the agent’s actions are beneficial to the principal. Also,

a main focus of the theory on the use of (financial) incentives corresponds well with the subject of this thesis, namely financial-risk sharing. A decision rights approach could have shed an interesting light on the different kind of organisational arrangements between third parties and GPs, but questions of integration and ownership of assets are beyond the scope of this thesis. Other considerations are that the use and the ownership of assets are less important issues in general practice than they are for instance in medical specialist care. Furthermore, there are hardly relation-specific assets. Regarding the question of authority, we assumed that the professional and the physicians' individual autonomy make it difficult for third parties to interfere in the patient-physician relationship in a way that differs from the ways in a market relationship. Hence we assumed that the professional relationship with GPs makes that the agency problems within an employment situation do not significantly differ from the agency problems within a market relationship. A final consideration was that GPs are often self-employed.

By applying agency theory to the several relationships between the patient, the GP, the third party and the so-called regulator, we constructed a theoretical framework that served as the theoretical background for the further analysis of the relationship between third parties and GPs. Firstly, we analysed what the agency function of third parties in health care involves. In order to provide the in health care familiar phenomenon 'managed care' with a proper theoretical framework and a clear classification, we then analysed whether managed care can be placed within the framework of agency theory. Next, we analysed in depth one specific managed-care technique third-party agents may employ within their relationships with GPs: financial-risk sharing. We analysed and discussed the rationale for such risk-sharing arrangements and the way they can be structured, and evaluated several examples of such arrangements in practice. Finally, we discussed how systems of financial-risk sharing should be structured.

9.2 Third-party agents and general practitioners

The health-care sector is characterised by uncertainty at the demand side, by an asymmetry of information between demanders and providers of care, and by the presence of external effects. These characteristics form a justification for the presence of a third party in health care, besides the patient (first party) and the provider of care (second party). The three main functions of a third party are:

- the insurance function, which consists mainly of pooling risks and paying claims;
- the agency function, which consists of buying care for a certain population, limiting moral hazard, and collecting and providing information about care;
- the access function, which is about guaranteeing the accessibility of, at least, some basic health-care goods and services.

In this study we focussed on one type of third party, namely a risk-bearing party that receives contributions for a defined group of members and for a certain period of time, and has the contractual or legal obligation to reimburse or to provide a specified package of health-care goods and services. More specific, we focussed on third parties that also act as agent on behalf of the insured. These third-party agents have to enter into relation-

ships with providers of care in order to fulfil their role properly. From their viewpoint, the relationship with GPs may be of interest then, as these physicians often act as gatekeepers and co-ordinators of medical services and have a considerable influence on the nature, the quantity, the quality and the costs of health care.

There are several ways the relationships between third parties and GPs can be structured. There are also several ways these relationships may be analysed and classified. We reviewed four classification schemes. Two major conclusions can be drawn then. Firstly, there are different approaches to the relationships, like a juridical, an organisational or a financial. Moreover, within a certain approach classification schemes focus on different aspects of the relationships. For instance, the option of an intermediary organisation (a middle tier) is sometimes overlooked. Secondly, the financial and the organisational sides of the relationships can differ considerably. These may vary from no direct relationship at all to full integration of both parties. Third parties may contract directly or indirectly (via a middle tier), pay physicians directly for the care provided or reimburse the insured, use several payment systems, et cetera. Further, a middle tier may translate a contract concluded with a third party into an altered contract with the physicians and change, for instance, the payment system.

9.3 Agency theory

Relationships between third parties and GPs are characterised by the presence of unequally distributed information and conflicting interests. Agency theory focuses on the relationship between two parties, specifically addresses the problems of asymmetric information and conflicting interests, and proposes strategies to deal with such problems (for instance, by suggesting optimal contract designs). In this theory, a principal commissions an agent to perform actions for, or on behalf of him. Their relationship is characterised by asymmetric information and conflicting interests. The agent has more information about his intentions, his actions or the circumstances upon which he bases his actions and may have other goals than the principal. As a result, problems of adverse selection and moral hazard may arise. Moreover, the principal may have problems drawing conclusions about the agent's efforts from the outcome, because in many agency relationships the outcome is uncertain due to external factors over which the agent has no control.

Some modifications of the traditional assumptions of the theory of agency have been proposed. These relate to the assumptions of goal conflict and opportunistic behaviour, the rationality of principal and agent, the programmability of the agent's tasks, the measurability of the outcome, the number of agents, and the duration of the relationship.

There are different strategies to handle the problems of agency relationships. In this thesis, we classified these strategies into three groups. Firstly, the principal may *select* an agent who is expected to behave in the principal's best interests. Then, the principal may try to *control* the agent. Different controlling strategies are control by incentives, control by persuasion or information, and control by directive or authority. By *monitoring* the agent, the principal may reduce the information gap and stimulate the agent to act in the principal's interests.

To accomplish desired behaviour, the principal has to incur monitoring and controlling costs. The agent may have to incur so-called bonding costs (expenses the agent has to incur when avoiding harming the principal or ensuring that the principal is compensated in case he acts contrary to the principal's best interests). Despite all the efforts, there still may be a residual loss due to the divergence between the realised outcome and the outcome the principal had in mind. Total agency costs are minimised by searching for a level of equilibrium between the monitoring and controlling plus bonding costs on the one hand and the residual loss on the other hand.

9.4 Agency and health care

Agency theory is not specifically aimed at relationships in health care, but in view of the characteristics of these relationships its use here seems justified. Regarding the relationship between patient and physician, some authors have even pointed at it as an obvious example of an agency relationship. Although less prominent, also the relationship between a third party and a physician has been described in agency terms.

Health-care relationships are not straightforward agency relationships, and in several respects they differ from the standard agency ones. Firstly, in the relationship with a GP the conflict of interests may be relatively small because of professional norms and values, social control by peers, et cetera that may guide him. Moreover, as a professional the physician may include at least part of the patient's interests in his own objectives. Secondly, the relationship between patient and GP usually is a lengthy or repeated one. This may stimulate the physician not to pursue solely his own interests. Poor performance may prompt the patient to terminate or not to renew the relationship and may also ruin the physician's reputation. Thirdly, due to his superior knowledge and experience, it will often be the agent (i.e. the physician) instead of the principal (i.e. the patient) who defines the goals that are attainable. Paradoxically, where in agency theory the information asymmetry is viewed as problematic, here the relatively ill-informed patient thus enters an agency relationship hoping for and trusting in the physician's superior knowledge and experience. Fourthly, the physician's informational advantage gives him the opportunity to induce the patient's demand for his goods or services. Although inducing demand is what a physician-as-agent should do, the physician may be able to persuade the patient to demand more goods or services than the patient probably would have demanded if he were as informed as his physician is. If the patient has health insurance then a special form of supplier-induced demand, namely supplier-induced moral hazard, may occur. Finally, the presence of a third party results in a triangular relationship. The third party may influence the information the patient and the physician possess as well as their interests (for instance, because of the altered financial incentives they face as a result of the presence of health insurance offered by the third party). Also, the presence and role of a third party has external effects since he has to take into account the preferences of other insured. Another effect of the presence of a third party is that it will be the third party, instead of the patient himself, who will make the contractual arrangements with the physician and try to promote that the physician acts in the patients best interests. This leaves

the patient as a principal with only a very limited set of techniques to influence the behaviour of the GP.

In spite of the fact that health-care relationships differ from standard agency relationships, the agency characteristics and the resulting agency problems are obviously present within the relationships described here. Hence agency theory is well applicable to relationships in the health care sector in general, and to the relationship between the third-party agent and the GP in particular.

We constructed a theoretical framework then that consists of four agency relationships and that provides the rationale for the use by a third party of strategies in order to influence the GPs.

The *first agency relationship* is between the patient (the principal) and the GP (the agent). Because of asymmetric information and diverging interests, the patient may encounter problems of adverse selection and moral hazard. In order to reduce the agency problems he may face within his relationship with the physician and to reduce his financial risk, the patient may want to enter into a relationship with a third-party agent; the *second agency relationship*. He may face difficulties then in selecting the third-party agent who will serve his interests best.

Once selected, the third-party agent is supposed to take measures to reduce the problems the patient may encounter within his relationship with the GP. To be an effective agent, the third party will have to enter into a relationship with the physician; the *third agency relationship*. Whether the third party is indeed an agent that serves the patient's best interests, is part of the problems within the second agency relationship.

The *fourth agency relationship* of the framework is the relationship between the third-party agent and the so-called regulator. In spite of the several techniques available to the insured, like (de)selection, control and monitoring, the insured's weak position creates a need for a regulator. The regulator should reduce the problems the insured may have within his relationship with the third-party agent and stimulate the latter to serve the insured's interests. A regulator may, for instance, further (managed) competition between third parties, provide information and institute supervision.

Within this framework, it is supposed that the insured as well as the regulator stimulate the third-party agent to gain information about the insured's preferences and to take measures in order to improve the agency relationship between patient and physician. Hence starting from standard agency theory, one might expect to find a third-party agent controlling and monitoring a selected group of physicians. For several reasons, like the presence of positive transaction costs and professional norms, the factual arrangements may differ from the theoretical arrangements, though. Moreover, it has been argued that relationships in health care may deviate from standard agency relationships. The question is therefore to what extent third-party agents actually use the three main strategies (selecting, controlling and monitoring) within their relationships with GPs.

9.5 Agency and managed care

The three main strategies derived from agency theory may result in the use of several techniques by the principal to further the outcome he aims for. Interestingly, we find such techniques within relationships between third-party agents and GPs as well. In health care (the use of) such a set of techniques is usually designated as 'managed care'. Although the term managed care is frequently used, a proper theoretical framework and a clear classification are lacking. We argued that the managed-care techniques fit in the triptych of agency theory remarkably well. We then pictured this triptych (selecting, controlling and monitoring) as three successive phases, together forming an iterative process: the agency cycle. Hence from the perspective of agency theory, managed care can be viewed as (the cyclical use of) a set of techniques by which the third-party agent attempts to influence the GP's behaviour in a way beneficial to the patient. We called this cyclical use the managed-care cycle.

Despite the difficulties associated with it, the selection and contracting of (primary care) physicians is regarded to be a crucial aspect of managed care. Some consider careful selection of physicians with a conservative practice style even to be the best guarantee of cost-effective and high-quality care. Research indicates that third parties prefer the selection of physicians before concluding contracts instead of 'pruning later'. The latter may be easier as more information about (the behaviour of) the physicians may be acquired, but it will be more difficult to change the physicians' behaviour by then or to get rid of them afterwards. However, selecting and then contracting the selected physicians is not simple. Selection requires sufficient and reliable qualitative information and quantitative data about the physicians' behaviour controlled for, for instance, practice size and case mix. Enhancing self-selection of physicians may reduce the selection problem, though. Once a subset of physicians is selected, they have to be contracted. Selective contracting requires an oversupply of physicians (in order to have a real choice), legal possibilities to contract selectively, the co-operation of individual physicians (the 'participation constraint') or of the profession as a whole and, last but not least, the consent of the insured.

The second phase of the managed-care cycle is controlling the physician. One prominent means to control the physician, the use of (financial) incentives, may help to stimulate the physician to choose from a set of possible actions the alternative that is most beneficial to the third party (and the patient as well). Financial incentives may emanate from the basic payment system and from ancillary payment systems. Predominant basic systems to pay GPs are fee for service, capitation and salary. Ancillary payments may be function-related, behaviour-related or outcome-related. Contrary to a function-related fee, behaviour- and outcome-related fees are made contingent upon an ex-post check on the way the specified services are provided and on the effects of the physician's behaviour respectively. A bonus system is one form of an ancillary payment. The third party may use mixed (blended) payment systems in order to balance the incentives from the several basic and ancillary payment systems. A mixed system may, for instance, balance the incentives for undesirable behaviour, like cream skimming and quality skimping, and

efficient behaviour as well as the incentives for quality efforts across contractible and noncontractible dimensions of quality.

Another way to control GPs by means of an incentive system is by their assignment to the gatekeeper function. The physician is stimulated to refer patients to other providers of care only if necessary. As it is a weak incentive, gatekeeping is often combined with other techniques.

Practice guidelines, physician or practice profiling and utilisation-management techniques are other prominent means to control physicians. Some of these techniques, like guidelines and profiling, aim at informing the physician and persuading him to perform the desired actions. Other techniques, like (pre-)admission review, mandatory second opinion and continued-stay review, aim at restricting the physician's choice. These techniques are sometimes viewed as an infringement on the professional autonomy or the individual physician's autonomy. However, an infringement on the individual physician's autonomy is exactly what is aimed at here and what is considered as a way to ensure that the patients' interests are served best. Whether such techniques are an infringement on the professional autonomy, depends on the role of the profession in designing and conducting the managed-care techniques. An association or a college of physicians, for instance, may issue practice guidelines. An example of practice guidelines are the eighty NHG-Practice Guidelines, developed by the Dutch College of General Practitioners (NHG). The same holds for monitoring physicians. Peer review is a widely accepted way of monitoring physicians by physicians themselves. In both examples the individual autonomy is in question but the professional autonomy is maintained. Because of this professional autonomy, physicians will prefer managed-care techniques designed or issued by the profession itself to techniques of a relative outsider, like a health insurer. Because of their individual autonomy, they will probably prefer control by incentives and control by persuasion or information to control by directives or authority. By means of directives or authority the physicians' actions are restricted in an almost coercive way. Choosing an alternative action is still an option, but at the risk of sanctions.

The final phase of the managed-care cycle is monitoring the physician. One goal is revealing information about the behaviour of the physician or the outcome of the process to which the physician contributed. In this way, the third party may try to reduce the informational advantage of the physician. A second goal is to reduce the information gap between physician and third party by informing the physician about his (relative) performance.

Managed care requires a relationship between third-party agent and physician. Within this relationship the several techniques may be used in combination with each other. The use of managed care gains in effectiveness if the three successive phases of the managed-care cycle, i.e. selection, control and monitoring, are designed and used coherently.

A problem for the third-party agent is that the outcome in terms of the patient's health status is uncertain and will only partially result from the GP's actions. It will result from the natural course of the disease, the medical treatment, the behaviour of the patient, and other health influencing factors. As a result, the third-party agent will have problems assessing the appropriateness of the physician's actions. Hence contracts that will specify

(all) possible outcomes and that, for instance, relate payments to factual outcomes are not common in health care. The majority of managed-care techniques focus on the behaviour of the physicians. Monitoring of physicians and analysing claim data may reveal important information then. Indeed, some third parties adjust their payments to physicians on the basis of utilisation, cost or quality measures, consumer surveys, physician productivity or other measures.

As the outcome in terms of health status will be the result of many factors, it is difficult to make the physician responsible for a negative outcome. The volume and the costs of the care provided, however, are closely related to physician's actions. Hence the third party may design a payment system with incentives that are directed at the volume or the costs of care. An example of such a system is financial-risk sharing, by which means the third party shares the responsibility for the costs of care with the physician. To safeguard the quality of care, the risk-sharing system has to be designed carefully and will additional controlling and monitoring techniques be necessary.

9.6 Financial-risk sharing in theory

The rationale for financial-risk sharing

A function of a third-party agent is to provide insurance against the insurance risk. This risk results from the incidence of illness, which has largely a stochastic nature. In return for an insurance premium, the risk is transferred from the insured to the third-party agent. This transfer involves a second risk, though, which is labelled here the risk of imperfect agency. It consists of the risk of the provision of cost-ineffective care, which mainly results from over-provision and inappropriate care, and of the risk of underprovision of care. The risk of imperfect agency may result from agency problems within the relationship between patient and physician, and from the presence of health insurance. Health insurance may lead to consumer-induced moral hazard as well as to supplier-induced moral hazard.

As the provision of health care is to large extent at the GP's discretion, the third-party agent may focus on this physician in order to reduce the size of the risk of imperfect agency. Once the third-party agent has taken over the insurance risk and the risk of imperfect agency from the insured, it may choose between four strategies to handle them. The first option is *risk bearing* in which the third party accepts the risk of imperfect agency as well as the insurance risk. This option is not very well compatible with the agency function, because the third party makes no attempt to influence the way health care is delivered. The second option is *risk shifting* in which the third party shifts both risks to individual physicians or to a group of physicians. The physicians become responsible for the insurance risk then, which however is typically a third-party function. The responsibility for both risks may prompt the physicians to take, from the third-party agent's viewpoint, undesirable measures to reduce their risk. An example of such undesirable behaviour is cream skimming. In the third option, *risk splitting*, the third party attempts to separate both risks after which it shifts the risk of imperfect agency to the physicians. Theoretically, this middle coarse between risk bearing and risk shifting may

be the most satisfactory solution. In practice, however, it will be very hard to separate both risks. The final option is *risk sharing*. Just like the third option, this option has the advantage of being a middle course between risk bearing and risk shifting. The difficulty of separating both risks is evaded though. A side-effect is that a part of the insurance risk is shifted to the (group of) physicians.

In a financial-risk sharing arrangement, the third-party agent stimulates the GP to reduce the amount of cost-ineffective care. Because the risk is shared, the incentives for over-provision and for underprovision can be balanced. Shifting a part of the risk (i.e. risk sharing) and the accompanying responsibilities has also the advantage that the decision-making is placed at a lower level (closer to the patient). The effectiveness of such an arrangement, however, depends on its specific financial and organisational design. The more the arrangement has the effect of the risk-shifting option, the stronger are both the incentives for cost-effective behaviour and for undesirable physician behaviour. There are at least three obvious forms of undesirable behaviour that a GP may show if he is at risk. In case of cream skimming the physician selects patients for whom the expected costs are lower than the reimbursement. In case of cost shifting the physician substitutes care for which he is not financially responsible for care for which he is. In case of quality skimping care is postponed or even withheld, efforts are reduced, et cetera.

The structure of risk-sharing arrangements

Financial-risk sharing arrangements have five main aspects that the third-party agent has to take into consideration while drawing up the arrangement. These aspects are the risk package, the size of the practice population, the normative level of care, the bonus system, and the limitation of the physician's risk.

A first crucial aspect is the scope of package of goods and services for which the physician is financially responsible, i.e. the design of the *risk package*. The type of care included will determine the probability that the physician incurs costs as well as the variability of the costs, given costs are made. For some types of care it will be easier to diagnose and to estimate the costs of treatment, which is especially important if the GP has to arrange and pay for follow-up care. Other matters to take into consideration are whether the risk package is divided into separate cost categories and whether the risk package influences the behaviour of other providers of care.

The second aspect is the *size of the practice population* or the proportion of it for which the physician is financially responsible. The relative and the absolute size of this population determine the magnitude of the incentives, the ability of the physician to shift costs to other parties and the extent that the physician is vulnerable to random fluctuations in the costs.

A third aspect is the *normative level of care*. The third party may define a norm, for instance a certain volume of care or a cost level, with which (the outcome of) the physician's behaviour is compared. This is probably the most difficult part of the arrangement. It is hard to determine an optimal level of care that is medically necessary and needs-based as well as cost-effective. An obvious way to determine a norm, then, is to use actual costs. In that case, a norm may be based on historical costs or on average costs. More sophisticated is econometric modelling to determine a norm that is adjusted for system-

atic differences in health status and for some of the other systematic factors, in so far as the physician can not influence them.

A fourth aspect is the *bonus system*. Eventually, the physician's financial responsibility may find expression in a bonus. A bonus is an ancillary payment paid if the physician has met certain requirements (like a financial norm). The negative variant of it is the malus, which the physician has to pay if he has not met the requirement. A bonus may be a fixed amount, may be proportional to the difference between actual and normative costs (the larger the difference, the larger the bonus) or may be inversely proportional to the difference between actual and normative costs (the smaller the difference, the larger the bonus). Other variants are systems in which the norm functions as a threshold (a bonus or a malus if the threshold is exceeded) or those in which the norm functions as a target (a bonus if the target is achieved).

The fifth aspect is the *limitation of the physician's risk* by means of additional measures. These measures are considered additional because the amount of risk is in the first instance the result of the risk package, the practice population, et cetera. The difficulty of separating the insurance risk from the risk of imperfect agency makes that the physician is also responsible for (a part of) the insurance risk. Without limiting this risk, there is a chance that the physician shows undesirable behaviour, that the incentive system will malfunction or that the physician is ruined. The incentive system may not function properly due to a few expensive patients in the first part of the financial year.

One way of risk reduction is reinsurance. A GP who has taken over part of the third party's risk may on his turn insure his liabilities. Although in the regular insurance industry reinsurance usually results in a risk contract with a second insurer (the reinsurer), in the present risk-sharing arrangements the third-party agent may function as a kind of reinsurer as well. Reinsurance is then a part of the financial arrangement between third party and physician. Not only may this result in lower costs, it also has the advantage that the third-party agent may balance the incentives from the reimbursement system with the incentives from the reinsurance system. Examples of reinsurance systems are a 'quota-share arrangement', an 'excess of loss per risk', 'an excess of loss per occurrence', or a 'stop loss'.

Another way to reduce the physician's risk is by means of risk pooling. In a risk-pooling arrangement a group of (primary care) physicians share together, possibly with other providers of care, in the rewards and penalties from surpluses and deficits in the budget(s) for a defined health-care package. Several variables determine the incentives emanating from the risk-pool arrangement, like the number of physicians or other providers, the number of patients, the proximity of the members of the pool (within the same building or scattered over a large region), et cetera. Other ways to vary the arrangements are by creating a multi-pool system and by adding one ore more intermediate organisations: the so-called middle tiers. This results in a myriad of options to allocate the financial risk.

9.7 Financial-risk sharing in practice

In some health care systems GPs are being put at risk for follow-up costs, like drug or hospital costs. This is no new phenomenon, since already in the first half of the twentieth century several third parties concluded risk contracts with physicians. Examples are the 'Zaanland system' and the 'Amsterdam system' in the Netherlands. Also the first Prepaid Group Practices in the United States are an example of early risk-sharing systems. Goal of the 'Zaanland system' and the 'Amsterdam system' was to reduce the costs of drugs in order to keep insurance premiums affordable and to enable a rise in the payments for medical specialists. The systems were indeed successful in curbing drug costs, but the restricted risk package stimulated the physicians to refer their patients. For several other reasons though, the systems were abolished during and after the Second World War. Another Dutch experience, the bonus-malus experiment in Tilburg during the 1980s, was also rather successful but was not followed by a permanent risk-sharing system.

Besides several Dutch systems, also British and North-American experiences with risk sharing were examined. In the United Kingdom, GPs could apply for fundholding status and receive a budget for follow-up care. They acted as purchasers of a health-care package on behalf of their practice population. Whether General Practice Fundholding was successful may always be up for discussion. The evidence suggests that fundholders managed to reduce the costs of some parts of their risk package, especially the costs of prescribing drugs. As the savings had to be used to improve patient care, it is likely that patients benefited from new services, quicker access to hospital care et cetera. As fundholders were smaller purchasers than Health Authorities, they were more flexible and were more able to shop around. The criticism that fundholding created a two-tier system indeed suggests that the system had positive effects and that fundholders' patients reaped the benefits of it. Whether this was at the expense of the patients of non-fundholders, remains a question to be answered. There is no evidence that the system provoked undesirable physician behaviour, like cream skimming or cost shifting.

Managed-care organisations in the US have a wide, but not always successful experience with risk sharing. Two cases were examined more in depth. Perhaps the most famous failure in the history of managed care is the United-Healthcare experience. United Healthcare contracted individual primary care physicians and shared the risks with them. The physicians had to act as gatekeepers and who had to control the volume and costs of their patients' health care. A very limited use of the large set of managed-care techniques and wrong choices in the design of the system finally resulted in the termination of the plan. The Blue Cross managed-care program Blue Plus seems to have been a more successful initiative. The financial incentives applied by Blue Plus, especially capitation but also fee for service with a withhold provision, seemed to reduce costs. Main difference with United Healthcare was that Blue Plus contracted medical group practices. These practices on their turn applied several managed-care techniques, including financial incentives linked to, for instance, individual physician productivity or the financial performance of the group.

The effects of different systems of financial-risk sharing
We described different risk-sharing arrangements here. The majority of these experiences at least indicate an effect of them on the behaviour of the GPs or primary care physicians. The 'Zaanland system' and the 'Amsterdam system' showed savings in the sickness funds' drug budgets. The 'Tilburg' experiment demonstrated a decrease in the number of referrals and hospitalisation days, as well as a slower increase in the number of physiotherapy treatments and in the costs of drugs. GP Fundholding appeared to be successful, for instance, in reducing (drug) costs and waiting times for specialist care, decreasing the number of hospital referrals and increasing the number of day case treatments. Blue Plus demonstrated reduced cost, both at the plan level and the physician level, although the effect differed per payment method. Less successful was United Healthcare, although costs per enrolee decreased spectacularly after structural financial and organisational changes were implemented. Whether this decrease resulted from altered physician behaviour or whether it, for instance, resulted from a selective disenrollment of high-cost insured is unclear.

Although the evidence indicates a differential effect of financial-risk sharing systems on the performance of GPs as agents for their patients, definite findings are especially hampered by methodological problems. A proper research design is often lacking. This makes it difficult to judge to what extent the physician's behaviour is affected by other (financial) incentives, by other managed-care techniques or by external factors. Further, studies of risk-sharing arrangements often provide little information on the quality of care, whereas this is an important outcome measure for judging the physicians' functioning as their patients' agents.

9.8 Towards a system of financial-risk sharing

Important lessons can be derived from the analysis of the theoretical and the practical models of financial-risk sharing. For several reasons these lessons are not definitive enough to devise a normative model of financial-risk sharing, though. Firstly, the practical examples that we analysed are limited in number. Hence the conclusions derived from these examples can not straightforwardly be applied to other health care systems. Further, econometric research is crucial in the design of a financial-risk sharing model. The several aspects of financial-risk sharing together determine the amount of risk that is transferred from third-party agent to GP and have to be balanced accurately. Finally, the ultimate form of financial-risk sharing will, for instance, differ per country and per health care system, will depend on the GPs' role and position, will depend on the preferences of the GP, and will be determined by the goals the third-party agent wants to achieve. Nevertheless, we have made a distinction between the design of an initial risk-sharing arrangement and the ultimate arrangement. The initial arrangement is characterised by:
- a limited risk package (for instance, drug costs only);
- a risk contract for at least fifty percent of the practice population;

- a norm (partly) based on historical figures to enable a relatively smooth transition to a risk-sharing system;
- a simple bonus system;
- an 'excess of loss per risk' in combination with a 'stop-loss contract', and a group practice or a locum group as risk pool.

A less conservative arrangement can be considered if a 'smooth' transition has been made to a risk-sharing system, third parties as well as physicians have gained experience with risk sharing and the effects on behaviour and outcomes are known. The arrangement may then be characterised by:

- a broad risk package from which expensive and 'open-ended treatments' are excluded;
- a risk contract preferably for the whole practice population;
- a norm that is adjusted for the patients' needs for care, that to some extent can predict future health-care costs, that is based on an optimal level of care and that functions as a target;
- a more sophisticated bonus system with a variable bonus, preferably an inversely proportional bonus system in combination with a norm as target;
- an 'excess of loss per risk' in combination with a 'stop-loss contract', and a group practice or a locum group as risk pool.

Regarding the bonus system, a third-party agent probably prefers an inversely proportional bonus with a norm as target as this makes explicit which goals should be achieved and as it makes adverse physician behaviour less likely. In case of a (inversely) proportional system, the bonus or malus should be substantial but less than hundred percent. Approximately fifty percent seems a good starting point then, but eventually a decision on the percentage will have to be taken in conjunction with the other aspects of financial-risk sharing.

It is crucial to balance all the aspects of a financial-risk sharing arrangement, the five main aspects and the additional aspects as well, very carefully. Furthermore, financial risk-sharing has to be accompanied by other techniques out of the managed-care cycle. Ideally, all the three phases of the managed-care cycle are represented in the design of the arrangement. Only then a risk-sharing arrangement may be acquired that helps to achieve the third-party agent's goals of efficient and high-quality care.

9.9 Epilogue

Although the behaviour of GPs partially determines the efficiency of the health-care sector, only little empirical and theoretical conceptual research has been done into the (effects of) financial incentives that may influence this behaviour. This holds especially for financial-risk sharing between third parties and GPs, but also for the broader set of managed-care techniques that the third party may use in an attempt to influence the behaviour of GPs. By constructing a conceptual framework for financial-risk sharing, by using this for the analysis of examples of risk-sharing systems, and by learning lessons for the structuring of such systems, we aimed at contributing to the literature on payment systems for GPs. By linking agency theory and managed care and by classifying the dif-

ferent techniques and bringing them in connection with each other by means of the managed-care cycle, we also aimed at contributing to the literature on managed care. These contributions are relevant because payment systems, financial-risk sharing and the other managed-care techniques are important issues in the worldwide discussion about improving the efficiency of health care. Based on the here described theoretical and empirical findings, we think there are good reasons to implement financial-risk sharing between third parties en GPs in those health-care systems in which the third party bears all the risks. For those systems though further econometric research may be necessary, and experimenting will have to prove whether the introduction of financial-risk sharing is attainable and desirable.

REFERENCES

Alchian, A.A. and H. Demsetz (1972), Production, information costs, and economic organization, American Economic Review 62, 777-795.

Arrow, K.J. (1963), Uncertainty and the welfare economics of medical care, American Economic Review LIII (5), 941-973.

Arrow, K.J. (1986), Agency and the market, in: K.J. Arrow and M.D. Intriligator (eds.), Handbook of mathematical economics III, Elsevier Science Publishers, Amsterdam, 1183-1195.

Atkinson, C. and A. Holbourn (eds.) (1994), Fundholding management handbook, Longman Group Limited, Harlow.

Audit Commission (1995), Briefing on GP fundholding, HMSO, London.

Audit Commission (1996), What the doctor ordered – A study of GP fundholders in England and Wales, HMSO, London.

Bailey, J.J. et al. (1994), Specialist outreach clinics in general practice, British Medical Journal 308, 1083-1086.

Bailit, H.L. and C. Sennett (1991), Utilization management as a cost-containment strategy, Health Care Financial Review 13, 87-93.

Bain, J. (1992), Budget holding in Calverton: one year on, British Medical Journal 304, 971-973.

Bain, J. (1994), Fundholding: a two tier system?, British Medical Journal 309, 396-399.

Bakker, F.M. (1997), Effecten van eigen betalingen op premies voor ziektekostenverzekeringen (Effects of cost sharing on premiums for health insurance), PhD thesis, Erasmus Universiteit Rotterdam, Rotterdam.

Begeleidingscommissie Uitvoering Geneesmiddelenbeleid (1999), Een helder recept – beleidsdoelen extramurale geneesmiddelenvoorziening vergen andere aanpak, Ministerie van Volksgezondheid, Welzijn en Sport, Den Haag.

Bindman, A.B. et al. (1998), Selection and exclusion of primary care physicians by managed care organizations, Journal of the American Medical Association 279 (9), 675-679.

Blomqvist, Å. (1991), The doctor as double agent: information asymmetry, health insurance, and medical care, Journal of Health Economics 10, 411-432.

Blox, J.T.H.M. et al. (1989), Bedrijfseconomie, H.E. Stenfert Kroese, Antwerpen.

Blumberg, L.J. and J. Holahan (2004), Government as reinsurer: potential impacts on public and private spending, Inquiry 41, 130-143.

Boerma, W.G.W. et al. (1997), Service profiles of general practitioners in Europe, British Journal of General Practice 47 (421), 481-486.

Boland, P. (1985), The role of preferred provider contracting in the healthcare market, in: P. Boland (ed.), The new healthcare market – A guide to PPO's for purchasers, payors and providers, Dow Jones-Irwin, Homewood, Illinois, 2-18.

Boston Consulting Group (1999), Geneesmiddel verzekerd – een nieuwe rol voor zorgverzekeraars in het inkoopproces van geneesmiddelen, advies aan Zorgverzekeraars Nederland.

Bowie, C. and R. Spurgeon (1994), Better data needed for analysis, British Medical Journal 309, 34.

Bradlow, J. and A. Coulter (1993), Effect of fundholding and indicative prescribing schemes on general practitioners' prescribing costs, British Medical Journal 307, 1186-1189.

Breedveld, E.J. et al. (1994), Ontwikkeling van een contracteringsmodel tussen huisarts en particuliere verzekeraar, instituut Beleid en Management Gezondheidszorg, Erasmus Universiteit Rotterdam.

Brennfleck Pascuzzi, E. (1993), Claims and benefits administration, in: P.R. Kongstvedt (ed.), The managed health care handbook, 2nd edition, Aspen Publishers, Gaithersburg, Maryland, 211-230.

Burr, A.J. et al. (1992), Impact of fundholding on general practice prescribing patterns, The Pharmaceutical Journal suppl. 249.

Carter, R.L. (1979), Reinsurance, Kluwer Publishers in association with Mercantile and General Reinsurance Company Ltd., Brentford.

Clark, D. and J.A. Olsen (1994), Agency in health care with an endogenous budget constraint, Journal of Health Economics 13, 231-251.

College Tarieven Gezondheidszorg (2004), brief aan de Minister van Volksgezondheid, Welzijn en Sport, kenmerk JDYK/mmor/A/04/114, met als onderwerp 'eerste rapportage nieuwe bekostiging huisartsen per 1 januari 2005 plus vastgestelde beleidsregel invoering consult- en visitetarief per 1 januari 2005 (ziekenfonds)', Utrecht, 30 september 2004.

Commissie Toekomstige Financieringsstructuur Huisartsenzorg (2001), Een gezonde spil in de zorg, Ministerie van Volksgezondheid, Welzijn en Sport, Den Haag.

Commissie Keuzen in de zorg (1991), Kiezen en delen – Advies in hoofdzaken, Ministerie van Welzijn, Volksgezondheid en Cultuur, Rijswijk.

Commissie modernisering curatieve zorg (1994), Gedeelde zorg: betere zorg, Commissie modernisering curatieve zorg.

Corney, R. (1994), Experiences of first wave general practice fundholders in South East Thames Regional Health Authority, British Journal of General Practice 44, 34-37.

Coulter, A. and J. Bradlow (1993), Effect of NHS reforms on general practitioners' referral patterns, British Medical Journal 306, 433-437.

Cromwell, J. and J.B. Mitchell (1986), Physician-induced demand for surgery, Journal of Health Economics 5, 293-313.

Crump, B.J. et al. (1991), Fundholding in general practice and financial risk, British Medical Journal 302, 1582-1584.

Crump, B.J. et al. (1995), Transferring the costs of expensive treatments from secondary to primary care, British Medical Journal 310, 509-512.

Culyer, A.J. (1989), The normative economics of health care finance and provision, Oxford Review of Economic Policy 5 (1), 34-58.

De Wit, G.W. (1994), Risico's, in: L.A.A. van den Berghe et al. (eds.), Heterogeniteit in verzekering: liber amicorum G.W. de Wit, Erasmus Insurance Center, Rotterdam, Instituut voor Actuariaat & Econometrie, Amsterdam.

Delnoy, D.M.J. and L.J. Stokx (1993), Huisarts: naast poortwachter nu ook rentmeester?, Medisch Contact 48 (24), 747-749.

Delnoy, D.M.J. et al. (1992), Geld, zorg en geldzorgen: honorering van huisartsen als instrument voor kostenbeheersing, Nederlands instituut voor onderzoek van de eerstelijnszorg, Utrecht.

Department of Health (1989), Working for patients, HMSO, London.

Dixon, J. et al. (1994), Distribution of NHS funds between fundholding and non-fundholding practices, British Medical Journal 309, 30-34.

Dowell, J.S. et al. (1995), Changing to generic formulary: how one fundholding practice reduces prescribing costs, British Medical Journal 310, 505-508.

Dranove, D. and W.D. White (1987), Agency and the organization of health care delivery, Inquiry 24 (Winter 1987), 405-415.

Drummond, M. et al. (1990), General practice fundholding, British Medical Journal 301, 1288-1289.
Eggleston, K. (2005), Multitasking and mixed systems for provider payment, Journal of Health Economics 24, 211-223.
Eisenberg, J.M. (2002), Physician utilization, Medical Care 40 (11), 1016-1035.
Eisenhardt, K.M. (1989), Agency theory: an assessment and review, Academy of Management Review 14 (1), 57-74.
Ellis, B.W. and H. Burns (1998), Why a new name for the Journal?, Journal of Integrated Care 2, 1-2.
Ellis, R. and T.G. McGuire (1990), Optimal payment systems for health services, Journal of Health Economics 9, 375-396.
Ellis, R.P. (1998), Creaming, skimping and dumping: provider competition on the intensive and extensive margins, Jounal of Health Economics 17, 537-555.
Enthoven, A.C. (1985), An economic analysis of the 'preferred provider organization' concept, in: P. Boland (ed.), The new healthcare market – A guide to PPO's for purchasers, payors and providers, Dow Jones-Irwin, Homewood, Illinois, 94-110.
Enthoven, A.C. (1994), On the ideal market structure for third-party purchasing of health care, Social Science & Medicine 39 (10), 1413-1424.
Evans, J.H. et al. (1995), Physicians' response to length-of-stay profiling, Medical Care 33 (11), 1106-1119.
Evans, R.G. (1974), Supplier-induced demand: some empirical evidence and implications, in: M. Perlman (ed.), The economics of health and medical care, MacMillan, London, 162-73.
Evans, R.G. (1981), Incomplete vertical integration: the distinctive structure of the health-care industry, in: J. van der Gaag and M. Perlman (eds.), Health, economics, and health economics, North-Holland Publishing Company, Amsterdam, 329-354.
Evans, R.G. (1984), Strained mercy: the economics of Canadian health care, Butterworths, Toronto.
Flierman, H.A. (1991), Changing the payment system of general practitioners, Nederlands instituut voor onderzoek van de eerstelijnsgezondheidszorg, PhD thesis, Rijksuniversiteit Utrecht, Utrecht.
Folland, S. et al. (1997), The economics of health and health care, 2nd edition, Prentice Hall, Upper Saddle River, New Jersey.
Forsberg, E. et al. (2001), Financial incentives in health care. The impact of performance-based reimbursement, Health Policy 58, 243-262.
Franks, P. et al. (1992), Gatekeeping revisited – protecting patients from overtreatment, New England Journal of Medicine 327, 424-429.
Fry, J. and J. Horder (1994), Primary health care in an international context, Nuffield Provincial Hospitals Trust, London.
Glaser, W.A. (1970), Paying the doctor, systems of remuneration and their effects, The Johns Hopkins Press, Baltimore and London.
Glennerster, H. (1998), Competition and budget devolution in health care: UK experience and future plans for change, Department of Social Policy, London School of Economics, London.
Glennerster, H. and M. Matsaganis (1993), The UK health reforms: the fundholding experiment, Health Policy 23, 179-191.
Glennerster, H. et al. (1992), A foothold for fundholding, King's Fund Institute, London.
Glennerster, H. et al. (1994), Implementing GP fundholding – wild card or winning hand?, Open University Press, Buckingham.
Glied, S. (2000), Managed care, in A.J. Culyer and J.P. Newhouse (eds.), Handbook of health economics, volume I, Elsevier Science, Amsterdam, 707-753.

Godber, E. et al. (1997), Economic evaluation and the shifting balance towards primary care: definitions, evidence and methodological issues, Health Economics 6 (3), 275-294.
Gogarty, M. and R. Halliday (1993), Effect of NHS reforms on GPs' referral patterns, British Medical Journal 306 , 716.
Gold, M.R. et al. (1995a), A national survey of the arrangements managed-care plans make with physicians, New England Journal of Medicine 333 (25), 1678-1683.
Gold, M.R. et al. (1995b), Behind the curve: a critical assessment of how little is known about arrangements between managed care plans and physicians, Medical Care Research and Review 52 (3), 307-341.
Gold, M.R. et al. (2002), Financial Risk Sharing with Providers in Health Maintenance Organizations, 1999, Inquiry 39, 34-44.
Gosden, T. et al. (1997), The efficiency of specialist outreach clinics in general practice: is further evaluation needed?, Journal of Health Services Research and Policy 2 (3), 174-179.
Gray, B.H. (1997), Trust and trustworthy care in the managed care era, Health Affairs 16, 34-49.
Grossman, S.J. and O.D. Hart (1986), The costs and benefits of ownership: a theory of vertical and lateral integration, The Journal of Political Economy 94 (4), 691-719.
Grytten, J. and R. Sørensen (2001), Type of contract and supplier-induced demand for primary physicians in Norway, Journal of Health Economics 20, 379-393.
Harris, C.M. and G. Scrivener (1996), Fundholders' prescribing costs: the first five years, British Medical Journal 313, 1531-1534.
Harris, M. and A. Raviv (1978), Some results on incentive contracts with applications to education and employment, health insurance, and law enforcement, American Economic Review 68 (1), 20-30.
Hart, O. and J. Moore (1990), Property rights and the nature of the firm, The Journal of Political Economy 98 (6), 1119-1158.
Hellinger, F.J. (1996), The impact of financial incentives on physician behavior in managed care plans: a review of the evidence, Medical Care Research and Review 53 (3), 294-314.
Hemenway, D. et al. (1990), Physicians' responses to financial incentives. Evidence from a for-profit ambulatory care center, New England Journal of Medicine 322, 1059-1063.
Hendrikse, G.W.J. (2003), Economics and management of organizations: co-ordination, motivation and strategy, McGraw-Hill, London.
Hendrikse G.W.J. and T. Jiang (2005), Plural form in franchising: an incomplete contracting approach, ERIM Report Series ERS-2005-090-ORG, Rotterdam.
Hickson, G.B. et al. (1987), Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study, Pediatrics 80 (3), 344-350.
Hillman, A.L. (1991), Managing the physician: rules versus incentives, Health Affairs, Winter 1991, 138-146.
Hillman, A.L. et al. (1992), Contractual arrangements between HMOs and primary care physicians: three-tiered HMOs and risk pools, Medical Care 30, 136-148.
Holmstrom, B. and P. Milgrom (1991), Multitask principal-agent analyses: incentive contracts, asset ownership, and job design, The Journal of Law, Economics & Organization 7 (special issue), 24-52.
Holmstrom, B. and P. Milgrom (1994), The firm as an incentive system, The American Economic Review 84 (4), 972-991.
Horgby, P.J. (1995), Risk management and health care, Cefos Report 5, Cefos, Göteborg.
Howie, J.G.R. et al. (1994), Evaluating care of patients reporting pain in fundholding practices, British Medical Journal 309, 705-710.

Howie, J.G.R. et al. (1995), Care of patients with selected health problems in fundholding practices in Scotland in 1990 and 1992: needs, process and outcome, British Journal of General Practice 45, 121-126.

Hsiao, W.C. (1992), Comparing health care systems: what nations can learn from one another, Journal of Health Politics, Policy and Law 17, 613-636.

Hurst, J.W. (1992), The reform of health care: a comparative analysis of seven OECD countries, OECD Health Policy Studies 2, Paris.

Iglehart, J.K. (1994), Health policy report – physicians and the growth of managed care, New England Journal of Medicine 331 (17), 1167-1171.

IOM (1989), Controlling costs and changing patient care? The role of utilization management, Report of a committee of the Institute of Medicine, Division of Health Care Services, National Academics Press, Washington.

Jack, W. (2005), Purchasing health care services from providers with unknown altruism, Journal of Health Economics 24, 73-93.

Janssen, R.T.J.D. (1988), Honorering van huisartsen - een verkenning van effecten op kosten en kwaliteit van de gezondheidszorg, Instituut voor Onderzoek van Overheidsuitgaven, Den Haag.

Jegers, M. et al. (2002), A typology for provider payment systems in health care, Health Policy 60, 255-273.

Jelovac, I. (2001), Physicians' payment contracts, treatment decisions and diagnosis accuracy, Health Economics, 10, 9-25.

Jensen, M.C. (1983), Organization theory and methodology, Accounting Review LVIII (2), 319-339.

Jensen, M.C. and W.H. Meckling (1976), Theory of the firm: managerial behavior, agency costs and ownership structure, Journal of Financial Economics 3, 305-360.

Johns, L. (1985), Case study: selective contracting in California, in: P. Boland (ed.), The new healthcare market – A guide to PPO's for purchasers, payors and providers, Dow Jones-Irwin, Homewood, Illinois, 948-965.

Kerr, E.A. et al. (1995), Managed care and capitation in California: how do physicians at financial risk control their own utilization?, Annals of Internal Medicine 123 (7), 500-504.

Kirkman-Liff, B.L. and W.P.M.M. van de Ven (1991), Capitated primary care physicians, preferred risk selection, and care-shifting: an international perspective, unpublished.

Kongstvedt, P.R. (1993a), Compensation of primary care physicians in open panels, in: P.R. Kongstvedt (ed.), The managed health care handbook, 2nd edition, Aspen Publishers, Gaithersburg, Maryland, 55-69.

Kongstvedt, P.R. (1993b), Changing provider behavior in managed care plans, in: P.R. Kongstvedt (ed.), The managed health care handbook, 2nd edition, Aspen Publishers, Gaithersburg, Maryland, 91-101.

Kongstvedt, P.R. (1993c), Formal physician performance evaluations, in: P.R. Kongstvedt (ed.), The managed health care handbook, 2nd edition, Aspen Publishers, Gaithersburg, Maryland, 189-198.

Kralewski, J.E. et al. (2000), The effects of medical group practice and physician payment methods on costs of care, Health Services Research 35 (3), 591-613.

Labelle, R. et al. (1994), A re-examination of the meaning and importance of supplier-induced demand, Journal of Health Economics 13, 347-368.

Langwell, K.M. (1990), Structure and performance of health maintenance organizations: a review, Health Care Financing Review 12 (1), 71-79.

Le Grand, J. (1991), Quasi-markets and social policy, Economic Journal 101, 1256-1267.

Lerner, C. and K. Claxton (1994), Modelling the behaviour of general practitioners, Discussion Paper 116, University of York, York.

Lindblom, C.E. (1977), Politics and markets, Basic Books, New York.
Luft, H.S. (1981), Health maintenance organizations: dimensions of performance, John Wiley & Sons, New York.
Luft, H.S. (1985), Foreword, in: D.P. Martin et al., A case study of united healthcare, The Henry J. Kaiser Family Foundation, Menlo Park, California.
Lurås, H. (2004), General practice: four empirical essays on GP behaviour and individuals preferences for GPs, Working Paper 2004: 1, University of Oslo, Oslo.
MacDonald, G.M. (1984), New directions in the economic theory of agency, Canadian Journal of Economics XVII (3), 415-440.
Majeed, A. and L. Malcolm (1999), Unified budgets for primary care groups, British Medical Journal 318, 772-776.
Martin, D.P. et al. (1985), A case study of united healthcare, The Henry J. Kaiser Family Foundation, Menlo Park, California.
Martin, S. et al. (1997), Risk and the GP budget holder, Discussion Paper 153, University of York, York.
Maturi, R.A. and Th.M. Raichel (1985), Preferred provider arrangements: market and delivery system perspectives, in: P. Boland (ed.), The new healthcare market – A guide to PPO's for purchasers, payors and providers, Dow Jones-Irwin, Homewood, Illinois, 19-37.
Max Geldens Stichting (1999), Op uw gezondheid!, Max Geldens Stichting, Amsterdam.
Maxwell, M. et al. (1993), General practice fundholding: observations on prescribing patterns and costs using the defined daily dose method, British Medical Journal 307, 1190-1194.
Maynard, A. (1994), Can competition enhance efficiency in health care?, Social Science & Medicine 39 (10), 1433-1445.
MDW-werkgroep Geneesmiddelen (1999), MDW-geneesmiddelen, Ministerie van Economische Zaken, Den Haag.
Miller, J.L. (1996), Distribution of capitation within physician organizations, in: P. Boland (ed.), The capitation sourcebook, Boland Healthcare, Berkeley, California, 63-95.
Miller, R.H. and H.S. Luft (1994), Managed care plan performance since 1980, Journal of the American Medical Association 271 (19), 1512-1519.
Miller, R.H. and H.S. Luft (2002), HMO plan performance update: an analysis of the literature, 1997-2001, Health Affairs 21 (4), 63-86.
Milstein, A. (1997), Managing utilization management: a purchaser's view, Health Affairs 16 (3), 87-90.
Minister van Volksgezondheid, Welzijn en Sport (2004), brief aan College Tarieven Gezondheidszorg, kenmerk woBOZ/PPB-2495136, met als onderwerp 'verzoek om uitvoeringstoets invoering nieuw bekostigingssysteem huisartsen', Den Haag, 7 juli 2004.
Mitnick, B.M. (1980), The political economy of regulation, Columbia University Press, New York.
Moe, T.M. (1984), The new economics of organization, American Journal of Political Science 28, 739-777.
Mooney, G. (1994), Key issues in health economics, Harvester Wheatsheaf, Hemel Hempstead.
Mooney, G. and M. Ryan (1993), Agency in health care: getting beyond first principles, Journal of Health Economics 12, 125-135.
Moore, S.H. et al. (1983), Does the primary-care gatekeeper control the costs of health care? Lessons from the SAFECO Experience, New England Journal of Medicine 309, 1400-1404.
National Audit Office (1994), Report by the comptroller and auditor general, General practitioner fundholding in England, HC 51 Session 1994-95, 9 December 1994.
Neelen, G.H.J.M. (1993), Principal-agent relations in non-profit organizations, Faculteit Bestuurskunde, Universiteit Twente, Enschede.

Newhouse, J.P. (1992), Medical care costs: how much welfare loss?, Journal of Economic Perspectives 6 (3), 3-21.
Newhouse, J.P. (1996), Reimbursing health plans and health providers: selection versus efficiency in production, Journal of Economic Literature 34, 1236-1263.
Newhouse, J.P. et al. (1996), Free for all?: lessons from the RAND Health Insurance Experiment, Harvard University Press, Cambrigde.
Newton, J. et al. (1993), Fundholding in Northern region: the first year, British Medical Journal 306, 375-378.
NHS Executive (1994), Developing NHS purchasing and GP fundholding, Department of Health, London.
NIVEL/RIVM (2004), Huisartsenzorg: wat doet de poortwachter?, Tweede nationale studie naar ziekten en verrichtingen in de huisartsenpraktijk 2, NIVEL/RIVM, Utrecht/Bilthoven.
NRV (1993), Health care in Europe – the finance and reimbursement systems of 18 European countries, Nationale raad voor de volksgezondheid, Zoetermeer.
Ohsfeldt, R.L. et al. (1998), The spread of state any willing provider laws, Health Services Research 33 (5), 1537-1562.
Ouchi, W.G. (1979), A conceptual framework for the design of organizational control mechanisms, Management Science 25 (9), 833-848.
Øvretveit, J. (1995), Purchasing for health, Open University Press, Buckingham.
Paritaire Werkgroep Huisartsenzorg (1995), Poortwachter in de praktijk, Paritaire Werkgroep Huisartsenzorg, Utrecht.
Parkerton, P.H. et al. (2003), Physician performance assessment. Nonequivalence of primary care measures, Medical Care 41 (9), 1034-1047.
Pauly, M.V. (1968), The economics of moral hazard: comment, American Economic Review 58, 531-537.
Pauly, M.V. (1978), Is medical care different?, in: W. Greenberg (ed.), Competition in the health care sector: past, present and future, Aspen Systems Corporation, Germantown Md., 11-35.
Pauly, M.V. (1986), Taxation, health insurance, and market failure in the medical economy, Journal of Economic Literature XXIV, 629-675.
Pauly, M.V. (1988a), Is medical care different? Old questions, new answers, Journal of Health Politics, Policy and Law 13 (2), 227-237.
Pauly, M.V. (1988b), Competition in health insurance markets, Law and Contemporary Problems 51 (2), 237-271.
Perrow, C. (1986), Complex organizations, 3rd edition, Random House, New York.
Petchey, R. (1995), General practitioner fundholding: weighing the evidence, Lancet 346, 1139-1142.
Phelps, C.E. (1992), Diffusion of information in medical care, Journal of Economic Perspectives 6 (3), 23-42.
Pontes, M.C. (1995), Agency theory: a framework for analyzing physician services, Health Care Manage Review 20 (4), 57-67.
Pratt, J.W. and R.J. Zeckhauser (1985), Principals and agents: an overview, in: J.W. Pratt and R.J. Zeckhauser (eds.), Principals and agents: the structure of business, Harvard Business School Press, Boston, Massachusetts.
President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research (1982), Making health care decisions, US Government Printing Office, Washington DC.
Propper, C. (1995a), Regulatory reform of the NHS internal market, Health Economics 4, 77-83.
Propper, C. (1995b), Agency and incentives in the NHS internal market, Social Science & Medicine 40 (12), 1683-1690.

Propper, C. et al. (1998), The effects of regulation and competition in the NHS internal market: the case of general practice fundholder prices, Journal of Health Economics 17, 645-673.
Raftery, J. and A. Stevens (1998), Day case surgery trends in England: the influences of target setting and of general practitioner fundholding, Journal of Health Services Research and Policy 3 (3), 149-152.
RCGP (1998), Factsheet GP Fundholding, Royal College of General Practitioners, London.
Reinhardt, U.E. (1985), The compensation of physicians: approaches used in foreign countries, Quality Review Bulletin 11, 366-377.
Remler, D.K. et al. (1997), What do managed care plans do to affect care? Results from a survey of physicians, Inquiry 34, 196-204.
Riley, K. (1997), The nuts and bolts of reinsurance, Lloyd's of London Press, London.
Robinson, J.C. (1993), Payment mechanisms, nonprice incentives, and organizational innovation in health care, Inquiry 30, 328-333.
Rodwin, M.A. (1995), Conflicts in managed care, New England Journal of Medicine 332, 604-607.
Ross, S.A. (1973), The economic theory of agency: the principal's problem, American Economic Review 63 (2), 134-139.
Rossiter, L.F. and G.R. Wilensky (1983), A reexamination of the use of physician services: the role of physician-initiated demand, Inquiry 20 (2), 62-172.
Ryan, M. (1994), Agency in health care: lessons for economists from sociologists, American Journal of Economics and Sociology 53 (2), 207-217.
Sappington, D.E.M. (1991), Incentives in principal-agent relationships, Journal of Economic Perspectives 5 (2), 45-66.
Schlesinger, M.J. et al. (1997), Medical professionalism under managed care: the pros and cons of utilization review, Health Affairs 16 (1), 106-124.
Schut, E. (1999), De markt voor gezondheidszorg, in: Ruud Lapré et al. (eds.), Algemene economie van de gezondheidszorg, Elsevier/de Tijdstroom, Maarssen.
Schut, F.T. (1986), Health Maintenance Organizations, De Tijdstroom, Lochem.
Schut, F.T. (1988), Verschillende markten binnen de sector gezondheidszorg, in: R.M. Lapré and F.F.H. Rutten (eds.), Economie van de gezondheidszorg, De Tijdstroom, Lochem, 197-221.
Schut, F.T. (1995), Competition in the Dutch health care sector, PhD thesis, Erasmus Universiteit Rotterdam, Rotterdam.
Scott, A. (1996), Agency, incentives and the behaviour of general practitioners: the relevance of principal agent theory in designing incentives for GPs in the UK, HERU Discussion Paper 03/96, Departments of Public Health and Economics, University of Aberdeen, Aberdeen.
Scott, A. (1997), Designing incentives for GPs. A review of the literature on their preferences for pecuniary and non-pecuniary job characteristics, HERU Discussion Paper 01/97, Departments of Public Health and Economics, University of Aberdeen, Aberdeen.
Scott, A. and J. Hall (1995), Evaluating the effects of GP remuneration: problems and prospects, Health Policy 31, 183-195.
Seal, H.L. (1969), Stochastic theory of a risk business, John Wiley & Sons, New York.
Shavell, S. (1979), Risk sharing and incentives in the principal and agent relationship, Bell Journal of Economics 10, 55-73.
Smith, P.C. et al. (1997), Principal-agent problems in health care systems: an international perspective, Health Policy 41, 37-60.
Sørensen, R.J. and J. Grytten (2003), Service production and contract choice in primary physician services, Health Policy 66, 73-93.
Spence, A.M. and R.J. Zeckhauser (1971), Insurance, information and individual action, American Economic Association, Papers and Proceedings, May 1971.

Spenceley, C. et al. (1994), NHS funds for fundholders and non-fundholders, British Medical Journal 309, 956.
Spremann, K. (1989), Agent and principal, in: G. Bamberg and K. Spremann (eds.), Agency theory, information, and incentives, Springer-Verlag, Berlin – Heidelberg.
Starfield, B. (1992), Primary care – concept, evaluation, and policy, Oxford University Press, New York.
Steiner, A. and R. Robinson (1998), Managed care: US research evidence and its lessons for the NHS, Journal of Health Services Research & Policy 3 (3), 173-184.
Stewart-Brown, S. et al. (1995), The effects of fundholding in general practice on prescribing habits three years after introduction of the scheme, British Medical Journal 311, 1543-1547.
STG (1997), Managed care en disease management in Nederland, Stichting Toekomstscenario's Gezondheidszorg, Zoetermeer.
Surender, R. et al. (1995), Prospective study of trends in referral patterns in fundholding and non-fundholding practices in the Oxford region, 1990-4, British Medical Journal 311, 1205-1208.
Tai-Seale, M. (2004), Voting with their feet: patient exit and intergroup differences in propensity for switching usual source of care, Journal of Health Politics, Policy and Law 29 (3), 501-511.
Tamblyn, R. et al. (2003), Physician and practice characteristics associated with the early utilization of new prescription drugs, Medical Care 41 (8), 895-908.
Tolley, H.D. et al. (1987), An evaluation of three payment strategies for capitation for medicare, Journal of Risk and Insurance 54, 678-690.
Toth, B. et al. (1997), Did the introduction of general practice fundholding change patterns of emergency admission to hospital?, Journal of Health Services Research and Policy 2 (1), 71-74.
Trapnell, G.R. (1985), Actuarial problems in PPOs, in: P. Boland (ed.), The new healthcare market – A guide to PPO's for purchasers, payors and providers, Dow Jones-Irwin, Homewood, Illinois, 285-297.
Tweede Kamer (2003-2004), Regeling van een sociale verzekering voor geneeskundige zorg ten behoeve van de gehele bevolking (Zorgverzekeringswet), Tweede Kamer, vergaderjaar 2003-2004, 29 763, nr. 3.
Van Barneveld, E.M. et al. (1998), Mandatory pooling as a supplement to risk-adjusted capitation payments in a competitive health insurance market, Social Science & Medicine 47 (2), 223-232.
Van de Ven, W.P.M.M. (1996), Market-oriented health care reforms: trends and future options, Social Science & Medicine 43, 656-666.
Van de Ven, W.P.M.M. and B. de Jong (1992), Huisarts als budgethouder, Inzet 16, III-IV.
Van de Ven, W.P.M.M. and R.C.J.A. van Vliet (1992), How can we prevent cream skimming in a competitive health insurance market? The great challenge for the 90's, in: P. Zweifel and H.E. Frech iii (eds.), Health economics wordlwide, Kluwer, 23-46.
Van de Ven, W.P.M.M. and R.P. Ellis (2000), Risk adjustment in competitive health plan markets, in: A.J. Culyer and J.P. Newhouse (eds.), Handbook of health economics I, Elsevier Science B.V., 755-845.
Van de Ven, W.P.M.M. et al. (1994), Forming and reforming the market for third-party purchasing of health care, Social Science & Medicine 39 (10), 1405-1412.
Van der Werf, W. (1997), Managed care – een nadere nuancering vanuit zorgaanbiederperspectief, Medisch Contact 52 (16), 511-513.
Van Doorslaer, E. (1988), Vraag naar gezondheidszorg, in: R.M. Lapré and F.F.H. Rutten (eds.), Economie van de gezondheidszorg, De Tijdstroom, Lochem, 67-88.
Van Doorslaer, E. and E. Schut (1999), Vraag naar gezondheidszorg, in: Ruud Lapré et al. (eds.), Algemene economie van de gezondheidszorg, Elsevier/de Tijdstroom, Maarssen.

Van Duuren, F. (1993), Het Zaanlandse en Amsterdamse stelsel – kostenbeheerssystemen in de jaren 30, Master thesis, Erasmus Universiteit Rotterdam, Rotterdam.
Van Tits, M.H.L. (1988), Vier jaar experimenteren met een bonus-malus-systeem onder huisartsen, IVA, Instituut voor Sociaal-Wetenschappelijk Onderzoek van de Katholieke Universiteit Brabant, Tilburg.
Van Tits, M.H.L. (1989), Experiment huisartsenhonorering, Medisch Contact 44 (8), 255-257.
Van Tits, M.H.L. and W.J.F.I. Nuyens (1987), Een bonus-malusexperiment onder huisartsen, Medisch Contact 42 (9), 276-279.
Vermaas, A. (1994), Financiële-risicodeling tussen zorgverzekeraars en huisartsen, Master thesis, Erasmus Universiteit Rotterdam, Rotterdam.
VNO-NCW (1995), Doelmatige gezondheidszorg: het werkt beter!, VNO-NCW, Den Haag.
Vogelaar, E. (2005), Voorstel van Ella Vogelaar, onafhankelijk voorzitter van het overleg tussen LHV, VWS en ZN, aan genoemde partijden voor een Beleidsagenda en bekostigingssystematiek huisartsenzorg voor 2006 en 2007, Ministerie van Volksgezondheid, Welzijn en Sport, Den Haag.
Von Eije, J.H. (1989), Reinsurance management: a financial exposition, PhD thesis, Erasmus Universiteit Rotterdam, Rotterdam.
Voûte, A.B.E. (1987), De waarde van verzekeren, in: J.H. von Eije et al. (eds.), Economie van het verzekeringsbedrijf, Kuwer, Deventer, 13-21.
Wagner, E.R. (1993), Types of managed care organizations, in: P.R. Kongstvedt (ed.), The managed health care handbook, 2nd edition, Aspen Publishers, Gaithersburg, Maryland, 12-21.
Walley, T. et al. (1995), Current prescribing in primary care in the UK, PharmacoEconomics 7 (4), 320-331.
Ward, D.L. (1993), Operational finance and budgetting, in: P.R. Kongstvedt (ed.), The managed health care handbook, 2nd edition, Aspen Publishers, Gaithersburg, Maryland, 281-298.
Weiner, J.P. and D.M. Ferriss (1990), GP budget holding in the United Kingdom: learning from American HMOs, Health Policy 16, 209-220.
Weiner, J.P. and G. de Lissovoy (1993), Razing a tower of Babel: a taxonomy for managed care and health insurance plans, Journal of Health Politics, Policy and Law 18 (1), 75-103.
Welch, H.G. et al. (1994), Physician profiling – an analysis of impatient practice patterns in Florida and Oregon, New England Journal of Medicine 330 (9), 607-612.
Welch, W.P. (1990), Giving physicians incentives to contain costs under Medicaid, Health Care Financing Review 12 (2), 103-112.
Welch, W.P. et al. (1990), Toward new typologies for HMOs, Milbank Quarterly 68 (2), 1990, 221-243.
Whynes, D.K. and G. Reed (1994), Fundholders' referral patterns and perceptions of service quality in hospital provision of elective general surgery, British Journal of General Practice 44, 557-560.
Whynes, D.K. et al. (1995), GP fundholding and the costs of prescribing, Journal of Public Health Medicine 17 (3), 323-329.
Whynes, D.K. et al. (1997), Prescribing cost savings by GP fundholders: long-term or short-term?, Health Economics 6, 209-211.
Williams, A. (1988), Priority setting in public and private health care, Journal of Health Economics 7, 173-183.
Williamson, O.E. (1985), The economic institutions of capitalism: firms, markets, relational contracting, Free Press, New York.
Wilson, R.P.H. et al. (1995), Alterations in prescribing by general practitioner fundholders: an observational study, British Medical Journal 311, 1347-1350.
Ziekenfondsraad (1993), Advies inzake gepast gebruik, Ziekenfondsraad, Amstelveen.

SAMENVATTING EN CONCLUSIE

1 Inleiding

'De vernieuwing van het verzekeringsstelsel zal er aan bijdragen dat de verzekeraars hogere eisen zullen gaan stellen aan hun contracten met zorgaanbieders, zowel kwalitatief als financieel. Verruiming van de contracteermogelijkheden voor verzekeraars en aanbieders, prestatiegerichte bekostigingssystemen (…) en het beschikbaar komen van vergelijkende informatie over prestaties van zorgaanbieders ondersteunen deze ontwikkeling' (Tweede Kamer 2003-2004, blz. 5). Deze passage uit de Memorie van toelichting bij het wetsvoorstel Zorgverzekeringswet schetst hoe de toenmalige Nederlandse regering de relatie tussen zorgverzekeraars en zorgaanbieders in het nieuwe verzekeringsstelsel voor zich zag. De Zorgverzekeringswet (ingevoerd per 1 januari 2006) geeft zorgverzekeraars meer invloed dan voorheen en draagt ertoe bij dat zij kunnen optreden als onderhandelingspartners van de zorgaanbieders. Door de in de Memorie van toelichting genoemde contracteermogelijkheden, zoals het (selectief) contracteren, het honoreren op basis van prestaties en het vergelijken van prestaties, kunnen zorgverzekeraars bijdragen aan het bereiken van de doelstellingen van het nieuwe zorgverzekeringsstelsel: meer doelmatigheid, minder centrale sturing en een goede toegankelijkheid.

Het onderscheid tussen ziekenfondsverzekering en particuliere verzekering is met de invoering van de Zorgverzekeringswet opgeheven. Aangezien in de financieringsstructuur voor de huisartsenzorg dit onderscheid nog steeds bestond, was een ingrijpende wijziging van deze financieringsstructuur noodzakelijk. Al meerdere keren, waaronder in 2001 door de Commissie toekomstige financieringsstructuur huisartsenzorg (de 'Commissie Tabaksblat'), was een uniforme financieringsstructuur geadviseerd. Het zou echter tot medio 2005 duren voordat de Landelijke Huisartsen Vereniging, het ministerie van VWS en Zorgverzekeraars Nederland het (met behulp van een bemiddelaar) eens werden over de wijze van uniformeren. Het belangrijkste punt van discussie vormde juist de bovengenoemde invloed die zorgverzekeraars over de beroepsgroep zouden krijgen. Zo was het oorspronkelijk de bedoeling dat zorgverzekeraars de zeggenschap kregen over een bedrag van € 138 miljoen, wat neerkwam op 20 procent van de praktijkkostenvergoeding die huisartsen tot dan toe kregen uitbetaald (ofwel circa 10 procent van de macrokosten van huisartsenzorg). Na protesten van de beroepsgroep werd dit bedrag vervolgens teruggebracht tot € 25 miljoen voor 2006, waarmee de invloed van de zorgverzekeraars weer werd beperkt.

Toch zijn er tussen de Landelijke Huisartsen Vereniging, het ministerie van VWS en Zorgverzekeraars Nederland enkele opmerkelijke afspraken gemaakt. De verplichting voor zorgverzekeraars om vrije beroepsbeoefenaren, waaronder huisartsen, te contracteren is al in 1992 afgeschaft. Doordat echter is afgesproken dat huisartsen voor het verkrijgen van een belangrijk deel van hun omzet een contract nodig hebben, lijkt er nu vanaf 2006 een omgekeerde contracteerplicht te zijn ingevoerd. Ten tweede zijn er af-

spraken gemaakt over het versterken van de poortwachters- en spilfunctie van de huisarts, over substitutie van tweede- naar eerstelijnszorg en over het voorschrijfbeleid van geneesmiddelen. De besparingen die deze afspraken moeten gaan opleveren zullen worden gebruikt voor de financiering van de huisartsenzorg. Er is daardoor sprake van een directe relatie tussen de prestaties van de beroepsgroep en de honorering ervan. Ten derde is afgesproken dat een stijging van het aantal contacten per patiënt per jaar met de huisartspraktijk zal leiden tot een verlaging van de consulttarieven. Omgekeerd zal een daling leiden tot een verhoging van de consulttarieven. Onduidelijk is voor hoeveel jaar deze afspraak zal gelden, maar voor de desbetreffende periode betekent deze in ieder geval dat het financiële risico voor zorgverzekeraars betreffende het volume van huisartsenzorg is afgewenteld op de beroepsgroep.

Het verschuiven van (een deel van) het financiële risico van zorgverzekeraars naar (een groep van) huisartsen is het onderwerp van dit onderzoek. Daarbij staan de volgende twee vragen centraal:

1. Zijn er argumenten voor financiële-risicodeling tussen zorgverzekeraars en huisartsen?
2. Zo ja, hoe zouden systemen waarin het financiële risico wordt gedeeld moeten worden vormgegeven?

Deze vragen zijn relevant omdat (financiële) prikkels in toenemende mate worden gezien als een effectieve manier om de doelmatigheid in de gezondheidszorg te vergroten. Een belangrijk onderdeel van hervormingen van de gezondheidszorg in het algemeen en van het zorgverzekeringsstelsel in het bijzonder is dat zogenaamde derde partijen, zoals publiek- of privaatrechtelijke zorgverzekeraars, worden geprikkeld om de efficiency en de kwaliteit van de gezondheidszorg te verbeteren. De bovenbeschreven invoering van de Zorgverzekeringswet is een voorbeeld van dergelijke hervormingen. Veelal worden zorgverzekeraars daarbij gedurende een bepaalde periode ten behoeve van een omschreven populatie verzekerden financieel verantwoordelijk gesteld voor een omschreven (basis-)verzekeringspakket. Op deze wijze worden zorgverzekeraars geprikkeld om door toepassing van diverse technieken de gestelde doelen van efficiency en kwaliteit te behalen. Contracteren van zorgaanbieders is dan een manier om invloed uit te oefenen op onder meer de wijze waarop en de prijs waartegen zorg wordt geleverd. Huisartsen zijn daarbij een interessante partij voor zorgverzekeraars, omdat zij door hun rol en positie een grote invloed hebben op de aard, het volume, de kwaliteit en de kosten van de zorg. Zorgverzekeraars kunnen in hun contracten met huisartsen afspraken maken over het gebruik van financiële prikkels, bijvoorbeeld door toepassing van financiële-risicodeling.

Het gedrag van artsen is mede bepalend voor de doelmatigheid van de gezondheidszorg. Gezien de wereldwijde belangstelling voor de doelmatigheid van de zorg, is het dan ook opmerkelijk hoe weinig betrouwbaar empirisch onderzoek er is naar het effect van financiële prikkels op het gedrag van artsen en op de uitkomsten van hun handelen. Er is ook betrekkelijk weinig theoretisch, conceptueel onderzoek voorhanden. Dit geldt vooral voor de honorering van huisartsen en voor financiële-risicodeling. Belangrijkste doelstelling van ons onderzoek is het creëren van een conceptueel raamwerk voor systemen van

financiële-risicodeling tussen zorgverzekeraars en huisartsen. Dit gebruiken we om verschillende praktijkvoorbeelden van financiële-risicodeling te analyseren. Hieruit zullen we vervolgens lessen trekken voor op te zetten systemen van risicodeling. We willen hiermee een bijdrage leveren aan de literatuur over honoreringssystemen voor huisartsen.

Financiële prikkels vormen echter slechts één mogelijkheid uit de set potentiële technieken die de zorgverzekeraar kan gebruiken om te trachten het gedrag van huisartsen te beïnvloeden. Deze set, of het gebruik ervan, wordt wel aangeduid met de term 'managed care'. Managed care is een diffuus begrip en mist een heldere theoretische inbedding. Voordat we een conceptueel raamwerk voor systemen van financiële-risicodeling tussen zorgverzekeraars en huisartsen creëren, achten we het dan ook noodzakelijk om eerst het ruimere fenomeen managed care van een theoretische inbedding en een heldere classificatie te voorzien. We willen hiermee een bijdrage leveren aan de literatuur over managed care.

In dit onderzoek hebben we gebruik gemaakt van agency-theorie om de relatie tussen zorgverzekeraars en huisartsen te analyseren. Om verschillende redenen hebben we voor een agency-perspectief gekozen. Een eerste reden is dat agency-theorie goed lijkt aan te sluiten bij de in dit proefschrift centraal staande situatie: de contractuele relatie tussen twee partijen (zorgverzekeraars en huisartsen) waarin de ene partij (de principaal) een relatie aangaat met een andere partij (de agent) in de verwachting dat de handelingen van de agent de principaal nut opleveren. Dat de theorie zich in belangrijke mate richt op het gebruik van (financiële) prikkels sluit goed aan bij het onderwerp van dit proefschrift, namelijk financiële-risicodeling. Een benadering waarin beslissingsrechten centraal staan zou een interessant licht hebben kunnen laten schijnen op de verschillende organisatorische wijzen waarop de relaties tussen zorgverzekeraars en huisartsen in de praktijk zijn vormgegeven. Vraagstukken van integratie en eigendom van activa, zoals (onroerende) goederen of apparatuur, vallen echter buiten het bereik van dit proefschrift. Andere overwegingen zijn dat het gebruik en het eigendom van dergelijke activa in de huisartsenzorg minder van belang zijn dan bijvoorbeeld in de medisch specialistische zorg. Bovendien is er in de huisartsenzorg nauwelijks sprake van relatie-specifieke activa. Wat betreft het vraagstuk van autoriteit, hebben we verondersteld dat de professionele autonomie en de autonomie van de individuele arts het voor een zorgverzekeraar moeilijk maken om zich te mengen in de relatie tussen arts en patiënt op een manier die afwijkt van de manier waarop de zorgverzekeraar dat in een marktrelatie kan. We hebben daarom verondersteld dat de professionele relatie met huisartsen ertoe leidt dat er voor wat betreft de agentschapproblemen geen wezenlijke verschillen zijn tussen situaties waarin huisartsen in dienstverband werkzaam zijn en situaties waarin huisartsen zelfstandig zijn gevestigd. Een laatste overweging betrof het feit dat huisartsen meestal zelfstandig zijn gevestigd.

Door agency-theorie eerst toe te passen op de relaties tussen de patiënt, de huisarts, de zorgverzekeraar en de overheid, hebben we een theoretisch raamwerk opgesteld. Dit heeft vervolgens gefungeerd als achtergrond voor de verdere analyse van de relatie tussen zorgverzekeraars en huisartsen. We hebben eerst geanalyseerd wat de agentschapsfunctie van een zorgverzekeraar inhoudt. Om het in de zorgsector bekende fenomeen 'managed care' van een goede theoretische inbedding en een heldere classificatie te voorzien, heb-

ben we vervolgens onderzocht of managed care kan worden geplaatst in het kader van de agency-theorie. Daarna zijn we nader ingegaan op één specifieke managed-caretechniek die zorgverzekeraars in hun relatie met huisartsen zouden kunnen toepassen: financiële-risicodeling. Na het analyseren en bediscussiëren van de argumenten voor financiële-risicodeling en de wijze waarop dergelijke systemen kunnen worden vormgegeven, hebben we enkele praktijkvoorbeelden van financiële-risicodeling geëvalueerd. Vervolgens zijn we ingegaan op de vraag hoe systemen met financiële-risicodeling zouden moeten worden vormgegeven. Ten slotte hebben we specifiek voor de Nederlandse situatie een eerste opzet gegeven voor experimenten waarin zorgverzekeraars en huisartsen de risico's delen.

2 De derde partij als agent en de huisarts

De gezondheidszorg wordt gekenmerkt door onzekerheid aan de vraagzijde van de markt, door een asymmetrische verdeling van informatie tussen vragers en aanbieders van zorg en door de aanwezigheid van externe effecten. Deze kenmerken vormen een rechtvaardiging voor de aanwezigheid, naast de patiënt (eerste partij) en de zorgaanbieder (tweede partij), van een derde partij in de gezondheidszorg. De drie hoofdfuncties van een derde partij zijn:

- de verzekeringsfunctie, voornamelijk bestaande uit het poolen van risico's en het betalen van claims;
- de agentschapsfunctie, bestaande uit de inkoop van zorg voor een bepaalde populatie, het terugdringen van 'moral hazard' en het verzamelen en leveren van informatie over de zorg;
- de toegangsfunctie, bestaande uit het garanderen van de toegankelijkheid van op zijn minst een basispakket gezondheidszorg.

In dit onderzoek hebben we ons gericht op één type derde partij, namelijk een risicodragende partij die gedurende een bepaalde periode premies ontvangt voor een omschreven groep verzekerden en die de contractuele of wettelijke plicht heeft om een omschreven pakket gezondheidszorggoederen en –diensten te leveren of te vergoeden. Meer specifiek hebben we ons gericht op derde partijen die ook als agent voor hun verzekerden optreden. Deze derde-partijen-als-agent, die we vanaf hier zorgverzekeraars zullen noemen, dienen relaties met zorgaanbieders aan te gaan om hun rol goed te kunnen vervullen. Vanuit hun standpunt bezien kunnen relaties met huisartsen zeer interessant zijn, omdat deze artsen vaak fungeren als poortwachters en regisseurs van medische zorg en daarmee een aanzienlijke invloed hebben op de aard, het volume, de kwaliteit en de kosten van de zorg.

De relaties tussen zorgverzekeraars en huisartsen kunnen op verschillende manieren worden vormgegeven. Er zijn ook verschillende manieren waarop deze relaties vervolgens kunnen worden geanalyseerd en geclassificeerd. Na bespreking van vier classificatiemodellen hebben we vervolgens twee conclusies getrokken. Ten eerste zijn er verschillende manieren om tegen de relaties tussen zorgverzekeraars en huisartsen aan te kijken, zoals een juridische, een organisatorische of een financiële benadering. Bovendien richten binnen een bepaalde benadering de modellen zich op verschillende aspecten van de

relaties. Zo wordt de mogelijke aanwezigheid van een tussenliggende organisatie (tussen zorgverzekeraar en huisarts) in het ene model als relevante factor beschouwd, maar in andere modellen buiten beschouwing gelaten. Ten tweede verschillen de relaties aanzienlijk wat betreft de financiële en organisatorische kanten ervan. Ze kunnen variëren van geen directe relatie tot volledige integratie van beide partijen. Verder kunnen zorgverzekeraars huisartsen direct of indirect (via een tussenliggende organisatie) contracteren, huisartsen direct betalen voor geleverde zorg of verzekerden de kosten ervan vergoeden, verschillende betalingssystemen hanteren enzovoorts. Vervolgens kan een tussenliggende organisatie een met de zorgverzekeraar gesloten contract vertalen in een afwijkend contract met huisartsen, bijvoorbeeld door de met de zorgverzekeraar overeengekomen betalingssystematiek te veranderen.

3 Agency-theorie

Relaties tussen zorgverzekeraars en huisartsen worden gekenmerkt door de aanwezigheid van ongelijk verdeelde informatie en conflicterende belangen. Agency-theorie richt zich op de relatie tussen twee partijen, houdt zich specifiek bezig met problemen als asymmetrische informatie en conflicterende belangen en stelt strategieën voor om die problemen te hanteren. In deze theorie schakelt een principaal een agent in om voor of namens hem bepaalde handelingen uit te voeren. Hun relatie wordt gekenmerkt door asymmetrische informatie en conflicterende belangen. De agent heeft meer informatie over zijn intenties, zijn handelen of de omstandigheden waarop hij zijn handelen baseert en heeft mogelijk andere doelen dan de principaal. Hierdoor kunnen problemen ontstaan als antiselectie en 'moral hazard'. Bovendien kan het voor de principaal moeilijk zijn om op basis van de uitkomst conclusies te trekken over de inspanningen van de agent, omdat de uitkomst veelal onzeker is als gevolg van externe omstandigheden waarop de agent geen invloed kan uitoefenen.

In de loop der tijd zijn verschillende aanpassingen van de traditionele aannames in de theorie voorgesteld. Deze hebben betrekking op de veronderstelde uiteenlopende doelen en opportunistische gedragingen, het rationele handelen van principaal en agent, de programmeerbaarheid van het handelen van de agent, de meetbaarheid van de uitkomst, het aantal agenten en de duur van de relatie.

Er zijn verschillende strategieën om de problemen van een agency-relatie te hanteren. We hebben deze strategieën in dit proefschrift ondergebracht in drie groepen. Ten eerste kan de principaal een agent *selecteren* waarvan hij veronderstelt dat deze zich in het belang van de principaal zal gedragen. Vervolgens kan de principaal trachten de agent te *sturen*, hetzij door prikkels, hetzij door overreding of informatie, hetzij door regels of macht. Ten derde kan de principaal door het *monitoren* van de agent trachten de informatieachterstand te verkleinen. Dit monitoren kan tevens dienen als motivatie voor de agent om in het belang van de principaal te handelen.

Om gewenst gedrag tot stand te brengen moet de principaal bepaalde kosten maken, bijvoorbeeld voor het sturen van de agent. Daarnaast zal de agent mogelijk ook bepaalde kosten maken, bijvoorbeeld om te voorkomen dat de principaal schade lijdt of om de principaal te kunnen compenseren als hij in strijd met het belang van de principaal han-

delt. Ondanks alle inspanningen kan de principaal nog steeds verlies lijden doordat de gerealiseerde uitkomst afwijkt van de uitkomst die de principaal in gedachten had. De totale agency-kosten worden geminimaliseerd door een evenwicht te zoeken tussen het beperken van dat verlies enerzijds en de kosten die daarvoor moeten worden gemaakt anderzijds.

4 Agency en gezondheidszorg

De agency-theorie is niet specifiek gericht op relaties in de gezondheidszorg, maar gelet op de kenmerken van deze relaties is het gebruik ervan in dit onderzoek alleszins gerechtvaardigd. De relatie tussen patiënt en arts wordt zelfs wel aangeduid als een typisch voorbeeld van een agency-relatie. Hoewel in mindere mate, is ook de relatie tussen een zorgverzekeraar en een (huis)arts wel in agency-termen beschreven.

Relaties in de gezondheidszorg zijn geen standaard agency-relaties en wijken daarvan in meerdere opzichten af. Ten eerste kunnen in de relaties met een huisarts de belangenconflicten relatief klein zijn, bijvoorbeeld doordat professionele normen en waarden en sociale controle door collega's richting geven aan zijn handelen. Bovendien maken voor de arts als professional de belangen van de patiënt mogelijk onderdeel uit van zijn eigen doelstellingen. Ten tweede is er tussen patiënt en huisarts vaak sprake van een langdurige of zich herhalende relatie, hetgeen de huisarts kan stimuleren niet slechts zijn eigen belangen na te streven. Ondermaatse prestaties van de huisarts kunnen voor de patiënt aanleiding zijn om de relatie te beëindigen of om de relatie te herzien en kunnen bovendien de reputatie van de huisarts schaden. Ten derde zal het door zijn kennis en ervaring vaak de agent (de huisarts) in plaats van de principaal (de patiënt) zijn die bepaalt welke doelen in hun relatie haalbaar zijn. Paradoxaal is dat waar in de agency-theorie de informatieasymmetrie als een probleem wordt gezien, de slecht geïnformeerde patiënt hier juist een agency-relatie start hopende en vertrouwende op de superieure kennis en ervaring van de arts. Ten vierde geeft de voorsprong in informatie van de arts hem de mogelijkheid om de vraag van de patiënt naar zijn goederen of diensten te beïnvloeden. Hoewel het beïnvloeden van deze vraag precies is wat de huisarts-als-agent moet doen, kan de huisarts de patiënt ervan overtuigen meer goederen of diensten te vragen dan de patiënt zou doen indien hij over dezelfde informatie zou beschikken als de arts. Als de patiënt een ziektekostenverzekering heeft kan bovendien een speciale vorm van aanbodgeïnduceerde vraag optreden, namelijk aanbodgeïnduceerde 'moral hazard'. Ten slotte resulteert de aanwezigheid van een zorgverzekeraar in een driehoeksrelatie. De zorgverzekeraar kan zowel de informatie waarover de patiënt en de arts beschikken als hun belangen beïnvloeden (bijvoorbeeld vanwege de veranderde financiële prikkels die zij ondervinden doordat de zorgverzekeraar ziektekostenverzekeringen levert). Ook ontstaan externe effecten doordat de zorgverzekeraar de voorkeuren van andere verzekerden in ogenschouw zal moeten nemen. Een ander effect van de aanwezigheid van een zorgverzekeraar is dat deze, in plaats van de patiënt zelf, de contractuele afspraken met de arts zal maken en zal trachten te bevorderen dat de arts handelt in het belang van de patiënt. Hierdoor beschikt de patiënt als principaal slechts over een beperkt aantal mogelijkheden om het gedrag van de huisarts te beïnvloeden.

Hoewel relaties in de gezondheidszorg veelal afwijken van de standaard agency-relaties, zijn de kenmerken van een agency-relatie en de resulterende problemen overduidelijk aanwezig. De agency-theorie is dan ook goed toepasbaar op relaties in de gezondheidszorg in het algemeen en op de relatie tussen de zorgverzekeraar en de huisarts in het bijzonder.

Vervolgens hebben we een theoretisch raamwerk opgesteld dat uit vier agency-relaties bestaat en waarmee wordt beargumenteerd waarom een zorgverzekeraar strategieën zou toepassen om het gedrag van huisartsen te beïnvloeden.

De *eerste agency-relatie* is de relatie tussen de patiënt (de principaal) en de huisarts (de agent). Door de asymmetrische informatie en de uiteenlopende belangen, kan de patiënt worden geconfronteerd met problemen als antiselectie en 'moral hazard'. Om zowel de agency-problemen die hij in de relatie met de arts ondervindt als zijn financiële risico te verminderen, kan de patiënt een relatie aangaan met de zorgverzekeraar; de *tweede agency-relatie*. Hij kan vervolgens problemen ondervinden bij het selecteren van een zorgverzekeraar die zijn belangen het best dient.

Zodra een zorgverzekeraar is geselecteerd wordt deze verondersteld maatregelen te treffen om de problemen die de patiënt ondervindt in zijn relatie met de huisarts terug te dringen. Om een effectieve agent te kunnen zijn, zal de zorgverzekeraar een relatie met de huisarts aan moeten gaan; de *derde agency-relatie*. Of de zorgverzekeraar ook daadwerkelijk als een agent in het belang van de patiënt handelt, is onderdeel van de problemen in de tweede agency-relatie.

De *vierde agency-relatie* in het raamwerk, is de relatie tussen de zorgverzekeraar en de zogenaamde 'spelregelbepaler'. Ondanks de mogelijkheden die een verzekerde tot zijn beschikking heeft (selecteren, sturen en monitoren), is de positie van de verzekerde dusdanig zwak dat er behoefte is aan bijvoorbeeld een overheidsinstantie die de 'spelregels' bepaalt. Deze overheidsinstantie moet de zorgverzekeraar stimuleren om in het belang van de verzekerde te handelen en zo de problemen beperken die deze verzekerde in zijn relatie met die zorgverzekeraar kan ondervinden. De overheid kan bijvoorbeeld (gereguleerde) concurrentie tussen zorgverzekeraars stimuleren, informatie verschaffen en toezicht instellen.

In het raamwerk wordt ervan uitgegaan dat zowel de verzekerde als de overheid de zorgverzekeraar stimuleert om informatie over de preferenties van de verzekerde in te winnen en om maatregelen te treffen teneinde de agency-relatie tussen patiënt en arts te verbeteren. Met de standaard agency-theorie in gedachten zou men dan verwachten een zorgverzekeraar aan te treffen die tracht het gedrag van een geselecteerde groep artsen te sturen en hun prestaties te monitoren. De feitelijke afspraken kunnen echter om verschillende redenen afwijken van de theoretische afspraken, zoals door de aanwezigheid van positieve transactiekosten en professionele normen. Bovendien wijken, zoals beargumenteerd, relaties in de gezondheidszorg af van de standaard agency-relaties. De vraag is dan ook in welke mate zorgverzekeraars als agent de drie strategieën (selecteren, sturen en monitoren) daadwerkelijk toepassen in hun relaties met huisartsen.

5 Agency en managed care

De drie strategieën die uit de agency-theorie naar voren zijn gekomen kunnen een principaal ertoe aanzetten om door toepassing van verschillende technieken de door hem gewenste uitkomst te bevorderen. Interessant is dat we dergelijke technieken ook aantreffen in de relaties tussen zorgverzekeraars en huisartsen. In de gezondheidszorg wordt (het gebruik van) een dergelijke set technieken doorgaans aangeduid met de term 'managed care'. Hoewel managed care een veel voorkomend begrip is, ontbreken een goede theoretische inbedding en een heldere classificatie ervan. We hebben hier beargumenteerd dat de managed-caretechnieken opmerkelijk goed passen in het drieluik van de agency-theorie. Vervolgens hebben we dit drieluik (selecteren, sturen en monitoren) voorgesteld als drie opeenvolgende fases die gezamenlijk een iteratief proces vormen: de agency-cyclus. Vanuit het perspectief van de agency-theorie kan managed care dan worden gezien als (het cyclische gebruik van) een set technieken waarmee de zorgverzekeraar tracht het gedrag van de huisarts zodanig te beïnvloeden dat deze in het belang van de patiënt handelt. Dit cyclische gebruik hebben we de managed-carecyclus genoemd.

Ondanks de moeilijkheden die ermee gepaard gaan, wordt het selecteren en contracteren van (huis)artsen gezien als een cruciaal onderdeel van managed care. Het zorgvuldig selecteren van artsen met een terughoudende manier van werken wordt zelfs wel beschouwd als de beste garantie voor kosteneffectieve zorg van hoge kwaliteit. Onderzoek geeft aan dat zorgverzekeraars er veelal de voorkeur aan geven om artsen te selecteren voordat contracten worden gesloten in plaats van achteraf maatregelen te moeten treffen. Voordeel van de laatste methode is weliswaar dat eenvoudiger informatie over (het gedrag van) de artsen kan worden verzameld, maar het zal dan moeilijker zijn om het gedrag van de artsen alsnog te veranderen of om nog van ze af te komen. Het selecteren en vervolgens contracteren van de geselecteerde artsen is echter niet eenvoudig. Selectie vergt voldoende en betrouwbare kwalitatieve en kwantitatieve informatie over het gedrag van de artsen die moet worden gecorrigeerd voor bijvoorbeeld praktijkomvang en samenstelling van de praktijkpopulatie. Het bevorderen van zelfselectie door artsen kan het selectieprobleem echter verkleinen. Zodra een groep artsen is geselecteerd, dienen ze te worden gecontracteerd. Selectief contracteren vergt een overschot aan artsen (wil er sprake kunnen zijn van een echte keuze). Verder dient selectief contracteren wettelijk te zijn toegestaan, dienen individuele artsen of dient de beroepsgroep als geheel bereid te zijn om mee te werken en, niet in de laatste plaats, dienen de verzekerden ermee in te stemmen.

De tweede fase van de managed-carecyclus is het sturen van de arts. Het gebruik van (financiële) prikkels is een belangrijke manier om te bevorderen dat de arts uit de verzameling mogelijke handelingen de handeling kiest die het meest nuttig is voor de zorgverzekeraar (en ook voor de patiënt). Financiële prikkels kunnen afkomstig zijn van zowel het basissysteem van honoreren als van allerlei aanvullende honoreringssystemen. De meest voorkomende basissystemen om huisartsen te betalen zijn betaling per verrichting, een abonnement of een salaris. Aanvullende betalingen kunnen gerelateerd zijn aan een functie, aan gedrag of aan een uitkomst. In tegenstelling tot een functiegerelateerde betaling, zijn gedrags- en uitkomstgerelateerde betalingen afhankelijk van respectievelijk een

controle achteraf op de wijze waarop bepaalde diensten zijn geleverd en op de effecten van het gedrag van de arts. Een bonussysteem is een vorm van aanvullende betalingen. De zorgverzekeraar kan gemengde honoreringssystemen toepassen om zo een evenwicht te vinden tussen de prikkels van de verscheidene basis- en aanvullende systemen. Een gemengd systeem kan bijvoorbeeld een evenwicht brengen in de prikkels voor ongewenst gedrag, zoals het selecteren van risico's en het bezuinigen op de kwaliteit van zorg, en doelmatig handelen. Ook kan een gemengd systeem een evenwicht brengen in de prikkels voor het leveren van inspanningen voor zowel contracteerbare als niet-contracteerbare dimensies van kwaliteit.

Een andere manier om huisartsen door prikkels te sturen is door toewijzing van de poortwachtersfunctie. De huisarts wordt op deze wijze gestimuleerd om patiënten alleen naar andere zorgaanbieders door te verwijzen indien dit noodzakelijk is. Aangezien hiervan slechts een zwakke prikkel uitgaat, wordt de poortwachtersfunctie vaak gecombineerd met andere technieken.

Praktijkstandaarden, het maken van profielen van huisartsen of van praktijken en zogenaamde 'utilisation-managementtechnieken' zijn andere in het oog springende manieren om artsen te sturen. Sommige technieken, zoals standaarden en profielen, hebben tot doel de arts te informeren en hem te overreden de gewenste handelingen uit te voeren. Andere technieken, zoals de verplichte second opinion en de zogenaamde '(pre)admission review' en 'continued-stay review', hebben tot doel de keuze van de arts te beperken. Dergelijke technieken worden wel beschouwd als een inbreuk op de professionele autonomie of de individuele autonomie van de arts. Een inbreuk op de individuele autonomie van de arts is juist precies wat hier wordt beoogd en wat wordt beschouwd als een manier om ervoor te zorgen dat de belangen van de patiënten het best wordt gediend. Of dergelijke technieken een inbreuk vormen op de professionele autonomie, hangt af van de rol van de beroepsgroep bij het ontwerpen en uitvoeren van de managed-caretechnieken. Zo kan een vereniging of een college van artsen standaarden uitgeven. Een voorbeeld van praktijkstandaarden zijn de inmiddels tachtig NHG-Standaarden die door het Nederlands Huisartsen Genootschap (NHG) zijn ontwikkeld. Hetzelfde geldt voor het monitoren van artsen. Intercollegiale toetsing is een algemeen geaccepteerde vorm van het monitoren van artsen door artsen zelf. In beide voorbeelden is de individuele autonomie in het geding, maar blijft de professionele autonomie gehandhaafd. Vanwege deze professionele autonomie zullen artsen de voorkeur geven aan managed-caretechnieken die zijn ontwikkeld of uitgegeven door de beroepsgroep zelf boven technieken van een relatieve buitenstaander, zoals een zorgverzekeraar. Vanwege hun individuele autonomie zullen ze waarschijnlijk de voorkeur geven aan sturing door prikkels en sturing door overreding of informatie boven sturing door regels of macht. Door regels of macht worden de handelingen van artsen immers op een nagenoeg dwingende manier beperkt. Het kiezen van een alternatieve handeling is weliswaar nog steeds mogelijk, maar met als risico dat sancties worden opgelegd.

De laatste fase van de managed-carecyclus is het monitoren van de arts. Eén doel ervan is het openbaren van informatie over het gedrag van de arts of de uitkomst van het proces waaraan de arts heeft bijgedragen. De zorgverzekeraar kan zo trachten zijn informatieachterstand ten opzichte van de arts te verkleinen. Een tweede doel is om de infor-

matiekloof tussen arts en zorgverzekeraar te overbruggen door juist de arts te informeren over zijn (relatieve) prestaties.

Managed care vereist een relatie tussen de zorgverzekeraar en de arts. In deze relatie kunnen vervolgens verschillende technieken in combinatie met elkaar worden gebruikt. De toepassing van managed care wint aan effectiviteit als de drie opeenvolgende fases uit de managed-carecyclus, te weten selecteren, sturen en monitoren, in samenhang worden vormgegeven en worden toegepast.

Een probleem voor de zorgverzekeraar is, dat wat betreft de gezondheidstoestand van de patiënt de uitkomst onzeker is en dat deze slechts gedeeltelijk het gevolg zal zijn van het handelen van de huisarts. De gezondheidstoestand zal de resultante zijn van het natuurlijke beloop van de ziekte, de medische behandeling, het gedrag van de patiënt en andere gezondheidsbeïnvloedende factoren. Hierdoor zal het voor de zorgverzekeraar lastig zijn om te beoordelen of de handelingen van de arts voldoen. Contracten waarin alle mogelijke uitkomsten zijn gespecificeerd en waarin bijvoorbeeld de vergoedingen zijn gerelateerd aan de feitelijke uitkomsten, zijn in de gezondheidszorg dan ook niet gebruikelijk. Van de managed-caretechnieken richt het merendeel zich op het gedrag van de artsen. Het observeren van artsen en het analyseren van claimdata kunnen daarbij belangrijke informatie opleveren. Zorgverzekeraars stemmen hun vergoedingen aan artsen dan af op basis van gebruik, kosten- of kwaliteitsmaatstaven, klantenonderzoeken, productiviteit van de arts of andere maatstaven.

Aangezien de uitkomst in termen van gezondheidstoestand door vele factoren zal worden bepaald, kan de arts moeilijk verantwoordelijk worden gesteld voor een negatieve uitkomst. Het volume en de kosten van de geleverde zorg zijn echter nauw aan het gedrag van de arts gerelateerd. De zorgverzekeraar kan daarom een systeem ontwerpen met prikkels die op het volume of de kosten van de zorg zijn gericht. Een voorbeeld van een dergelijk systeem is financiële-risicodeling, waardoor de zorgverzekeraar de verantwoordelijkheid voor de kosten van de zorg deelt met de arts. Om de kwaliteit van zorg te waarborgen, dient een systeem van risicodeling zorgvuldig te worden vormgegeven en zullen aanvullende sturings- en monitoringstechnieken nodig zijn.

6 Financiële-risicodeling in theorie

Argumenten voor financiële-risicodeling

Een zorgverzekeraar heeft onder meer als rol om verzekeringen aan te bieden tegen het verzekeringsrisico. Dit risico ontstaat doordat het vóórkomen van ziekten een grotendeels stochastisch karakter heeft. In ruil voor een verzekeringspremie wordt het risico van de verzekerde overgedragen aan de zorgverzekeraar. Deze overdracht betreft echter ook een tweede risico, dat wij het risico van imperfect agentschap hebben genoemd. Het bestaat uit het risico van ondoelmatige zorg, dat vooral ontstaat door overconsumptie en niet-gepaste zorg, en uit het risico van onderconsumptie. Het risico van imperfect agentschap kan het gevolg zijn van agency-problemen in de relatie tussen patiënt en arts en van de aanwezigheid van ziektekostenverzekeringen. Ziektekostenverzekeringen kunnen leiden

tot zowel consument-geïnduceerde moral hazard als aanbieder-geïnduceerde moral hazard.

Aangezien huisartsen bij de levering van gezondheidszorg voor een belangrijk deel naar eigen goeddunken kunnen handelen, kan de zorgverzekeraar zich op deze arts richten teneinde het risico van imperfect agentschap te beperken. Nadat de zorgverzekeraar het verzekeringsrisico en het risico van imperfect agentschap heeft overgenomen van de verzekerde, kan hij uit vier strategieën kiezen om deze te hanteren. De eerste optie is *het dragen van beide risico's*, waarbij de zorgverzekeraar zowel het verzekeringsrisico als het risico van imperfect agentschap accepteert. Deze optie is niet goed verenigbaar met de agentschapsfunctie, omdat de zorgverzekeraar geen pogingen onderneemt om de manier waarop gezondheidszorg wordt geleverd te beïnvloeden. De tweede optie is *het doorschuiven van beide risico's*, waarbij de zorgverzekeraar beide risico's naar individuele artsen of naar een groep van artsen doorschuift. De artsen worden dan verantwoordelijk voor het verzekeringsrisico, wat echter een typische functie voor een zorgverzekeraar is. De verantwoordelijkheid voor beide risico's kan de artsen ertoe aanzetten om, bezien vanuit het gezichtspunt van de zorgverzekeraar, ongewenste maatregelen te nemen om zo hun risico te beperken. Een voorbeeld van dergelijk ongewenst gedrag is selectie van gunstige risico's. In de derde optie, *het opsplitsen van beide risico's*, tracht de zorgverzekeraar beide risico's te scheiden, waarna hij het risico van imperfect agentschap doorschuift naar de artsen. Theoretisch gezien lijkt deze middenweg tussen het dragen van beide risico's en het doorschuiven van beide risico's de meest bevredigende oplossing. In de praktijk zal het echter moeilijk zijn om beide risico's van elkaar te scheiden. De laatste optie is *het delen van beide risico's*. Net als de derde optie heeft deze optie als voordeel dat er een middenweg is gevonden tussen het dragen van beide risico's en het doorschuiven van beide risico's. Het problematische scheiden van beide risico's wordt hierdoor echter vermeden. Bijkomend effect is wel dat een deel van het verzekeringsrisico naar (een groep van) artsen wordt doorgeschoven.

Door een regeling met financiële-risicodeling stimuleert de zorgverzekeraar de huisarts om de hoeveelheid ondoelmatige zorg te verminderen. Doordat het risico wordt gedeeld, kunnen de prikkels voor het leveren van te veel of voor het leveren van te weinig zorg in evenwicht worden gehouden. Het verschuiven van een deel van het risico (risicodeling) en de bijbehorende verantwoordelijkheden heeft bovendien als voordeel dat besluitvorming op een lager niveau (dichter bij de patiënt) komt te liggen. De effectiviteit van een dergelijke regeling wordt echter bepaald door de specifieke financiële en organisatorische vormgeving ervan. Naarmate de regeling meer de vorm aanneemt van de optie waarin beide risico's worden doorgeschoven, nemen de prikkels voor zowel doelmatig gedrag als voor ongewenst gedrag voor de arts toe. Een huisarts die risico loopt kan minstens drie voor de hand liggende vormen van ongewenst gedrag vertonen. Bij selectie van gunstige risico's selecteert de arts patiënten waarvan hij verwacht dat de kosten lager zullen uitvallen dan de vergoedingen. Bij het doorschuiven van de kosten substitueert de arts zorg waarvoor hij niet financieel verantwoordelijk is voor zorg waarvoor hij dat wel is. Bij het bezuinigen op de kwaliteit van zorg wordt zorg uitgesteld of zelfs onthouden, worden inspanningen verminderd enzovoorts.

De structuur van regelingen met financiële-risicodeling

Als een zorgverzekeraar een regeling met financiële-risicodeling opstelt, dan dient hij vijf hoofdaspecten in beschouwing te nemen. Deze aspecten zijn het risicopakket, de omvang van de praktijkpopulatie, het normatieve niveau van zorg, het bonussysteem en de beperking van het risico van de arts.

Een eerste cruciaal aspect is de omvang van het pakket goederen en diensten waarvoor de arts financieel verantwoordelijk is, oftewel de vormgeving van het *risicopakket*. Het type zorg dat wordt opgenomen bepaalt de kans dat de arts kosten moet maken alsook, gegeven dat kosten worden gemaakt, de variabiliteit van de kosten. Bij bepaalde types zorg zal het makkelijker zijn om een diagnose te stellen en om de kosten van behandeling in te schatten, wat vooral belangrijk is als de huisarts verantwoordelijk is voor het regelen en vergoeden van vervolgzorg. Andere punten die hierbij in ogenschouw moeten worden genomen zijn of het risicopakket wordt opgedeeld in verschillende kostencategorieën en of het risicopakket het gedrag van andere zorgaanbieders beïnvloedt.

Het tweede aspect is de *omvang van de praktijkpopulatie* of het deel daarvan waarvoor de arts financieel verantwoordelijk is. De relatieve en absolute omvang van deze populatie bepalen de grootte van de prikkels, de mogelijkheid voor de arts om kosten door te schuiven naar andere partijen en de mate waarin de arts kwetsbaar is voor toevallige schommelingen in de kosten.

Een derde aspect is het *normatieve niveau van zorg*. De zorgverzekeraar kan een norm definiëren, bijvoorbeeld een bepaald volume van zorg of een kostenniveau, waarmee (de uitkomst van) het gedrag van de arts wordt vergeleken. Dit is vermoedelijk het meest lastige onderdeel van de regeling. Het is moeilijk om een optimaal niveau te bepalen van zorg die zowel medisch noodzakelijk en op basis van behoefte is als doelmatig. Een norm kan in dat geval gebaseerd zijn op historische of gemiddelde kosten. Econometrisch modelleren is een geavanceerdere manier om een norm vast te stellen die is gebaseerd op systematische verschillen in gezondheidstoestand en enkele andere systematische factoren, voor zover de arts ze niet kan beïnvloeden.

Een vierde aspect is het *bonussysteem*. Uiteindelijk kan de financiële verantwoordelijkheid van de arts tot uitdrukking komen in een bonus. Een bonus is een aanvullende betaling die wordt betaald als de arts aan bepaalde vereisten heeft voldaan (zoals een financiële norm). De negatieve variant ervan is een malus, die de arts moet betalen als hij niet aan de vereisten heeft voldaan. Een bonus kan bestaan uit een vast bedrag, kan evenredig zijn aan het verschil tussen de werkelijke en de normatieve kosten (hoe groter het verschil, hoe groter de bonus) of kan omgekeerd evenredig zijn aan het verschil tussen de werkelijke en de normatieve kosten (hoe kleiner het verschil, hoe groter de bonus). Andere varianten zijn systemen waarin de norm fungeert als een drempel (een bonus of een malus als de drempel is overschreden) of die waarin de norm fungeert als een doel (een bonus als het doel is behaald).

Het vijfde aspect is het door middel van aanvullende maatregelen *beperken van het risico van de arts*. Deze maatregelen kunnen als additioneel worden beschouwd omdat de hoeveelheid risico in eerste instantie wordt bepaald door het risicopakket, de praktijkpopulatie enzovoorts. De moeilijkheid om het verzekeringsrisico en het risico van imperfect

agentschap van elkaar te scheiden leidt ertoe dat de arts ook verantwoordelijk wordt voor (een deel van) het verzekeringsrisico. Als dit risico niet wordt beperkt, bestaat de kans dat de arts ongewenst gedrag gaat vertonen, dat het systeem van prikkels niet meer goed functioneert of dat de arts wordt geruïneerd. Dat het systeem van prikkels niet meer goed functioneert, kan worden veroorzaakt door het hebben van een paar kostbare patiënten aan het begin van het financiële jaar.

Eén manier om het risico te beperken is door middel van herverzekering. Een huisarts die een deel van het risico van de zorgverzekeraar heeft overgenomen kan zich op zijn beurt verzekeren voor zijn aansprakelijkheden. Hoewel in de reguliere verzekeringswereld herverzekering doorgaans resulteert in een contract met een tweede verzekeraar (de herverzekeraar), kan in de onderhavige risicodelingsregelingen de zorgverzekeraar eveneens als een soort herverzekeraar optreden. Herverzekering vormt dan een onderdeel van de regeling tussen zorgverzekeraar en arts. Niet alleen kan dit tot lagere kosten leiden, het heeft ook als voordeel dat de zorgverzekeraar de prikkels die van het systeem van vergoedingen uitgaan en de prikkels die van het herverzekeringssysteem uitgaan met elkaar in evenwicht kan brengen. Voorbeelden van herverzekeringssystemen zijn een 'quota-share' regeling, een 'excess of loss per risk', een 'excess of loss per occurrence' of een 'stop loss'.

Een andere manier om het risico van de arts te beperken is door middel van risicopooling. In een regeling met risicopooling deelt een groep (huis)artsen gezamenlijk, eventueel met andere aanbieders van zorg, in de beloningen en boetes die volgen uit overschotten of tekorten in het budget of in de budgetten voor een omschreven gezondheidszorgpakket. Verschillende variabelen bepalen de prikkels die uitgaan van een regeling met risicopooling, zoals het aantal artsen of andere zorgaanbieders, het aantal patiënten, de nabijheid van de leden van de pool (in hetzelfde gebouw of verspreid over een grote regio) enzovoorts. Andere manieren om de regelingen te variëren zijn het vormen van een meervoudige pool en het toevoegen van één of meerdere tussenliggende organisaties. Dit resulteert in een grote hoeveelheid opties voor de verdeling van het financiële risico.

7 Financiële-risicodeling in praktijk

In enkele gezondheidszorgsystemen dragen huisartsen risico voor vervolgkosten, zoals geneesmiddelen of ziekenhuiskosten. Dit is geen nieuw fenomeen, aangezien verschillende zorgverzekeraars al in de eerste helft van de twintigste eeuw risicocontracten met artsen afsloten. Voorbeelden zijn het Zaanlandse stelsel en het Amsterdamse stelsel in Nederland. Ook de eerste 'Prepaid Group Practices' in de Verenigde Staten zijn een voorbeeld van vroege systemen met risicodeling. Doel van het Zaanlandse stelsel en het Amsterdamse stelsel was de kosten van geneesmiddelen te beperken om zo de verzekeringspremies betaalbaar te houden en een stijging van de betalingen voor medisch specialisten mogelijk te maken. De stelsels waren inderdaad succesvol in het beteugelen van de kosten van geneesmiddelen, maar het beperkte risicopakket stimuleerde de artsen om hun patiënten te verwijzen. Om verscheidene andere redenen echter, werden de stelsels tijdens en na de Tweede Wereldoorlog opgeheven. In Nederland is verder gedurende de

jaren tachtig ervaring opgedaan met het bonus-malusexperiment in Tilburg. Ook dit was succesvol, maar heeft niet geleid tot een permanent systeem met risicodeling.

Behalve de Nederlandse systemen zijn ook Britse en Noord-Amerikaanse ervaringen met risicodeling bekeken. In het Verenigd Koninkrijk konden huisartsen zich opgeven voor budgethouderschap en een budget ontvangen voor vervolgzorg. Ze traden namens hun praktijkpopulatie op als zorginkopers. Of 'GP Fundholding' succesvol was zal wel altijd onderwerp van discussie blijven. De onderzoeksresultaten lijken uit te wijzen dat budgethouders in staat waren om de kosten voor bepaalde onderdelen van hun risicopakket te beperken, vooral de kosten van geneesmiddelen. Omdat de besparingen moesten worden gebruikt om de patiëntenzorg te verbeteren, is het waarschijnlijk dat patiënten baat hebben gehad bij nieuwe diensten, snellere toegang tot ziekenhuiszorg enzovoorts. Doordat budgethouders kleinere zorginkopers waren dan de 'Health Authorities' waren ze flexibeler en beter in staat om zorg elders in te kopen. De kritiek dat budgethouderschap tweedeling creëerde geeft aan dat het systeem inderdaad positieve effecten had en dat de patiënten van budgethouders er de vruchten van plukten. Of dit ten koste ging van de patiënten van niet-budgethouders blijft onduidelijk. Er is geen bewijs dat het systeem ongewenst gedrag van artsen, zoals het selecteren van gunstige risico's of het doorschuiven van de kosten, uitlokte.

Managed-careorganisaties in de Verenigde Staten hebben een ruime, doch niet altijd succesvolle ervaring met risicodeling. Twee voorbeelden zijn nader bestudeerd. Wellicht het meest beroemde fiasco uit de geschiedenis van managed care is dat van United Healthcare. United Healthcare contracteerde individuele huisartsen en deelde de risico's met hen. De artsen fungeerden als poortwachters en dienden het volume en de kosten van de zorg voor hun patiënten te beheersen. Een zeer beperkt gebruik van managed-caretechnieken en verkeerde keuzes in het ontwerp van het systeem leidden uiteindelijk tot de beëindiging van het verzekeringsmodel. Het managed-careprogramma Blue Plus van Blue Cross lijkt een succesvoller initiatief te zijn geweest. De financiële prikkels die door Blue Plus werden toegepast, met name het abonnementssysteem maar ook betaling per verrichting met een voorziening waarin vooraf geld werd ingehouden (een zogenaamde 'withhold'), leken de kosten te beperken. Het belangrijkste verschil met United Healthcare was dat Blue Plus groepspraktijken contracteerde. Deze praktijken pasten op hun beurt verscheidene managed-caretechnieken toe, waaronder financiële prikkels die bijvoorbeeld waren gekoppeld aan de productiviteit van de individuele arts of aan de financiële prestaties van de groep.

De effecten van verschillende systemen met financiële-risicodeling

We hebben hier verschillende systemen met financiële-risicodeling beschreven. De meerderheid van de ervaringen die hiermee zijn opgedaan wijzen minimaal op een effect ervan op het gedrag van de huisartsen. Het Zaanlandse stelsel en het Amsterdamse stelsel lieten besparingen zien in de ziekenfondsenbudgetten voor geneesmiddelen. Het Tilburgse experiment vertoonde zowel een daling in het aantal verwijzingen en ziekenhuisdagen als een lagere stijging van het aantal fysiotherapeutische behandelingen en van de kosten van geneesmiddelen. Het Britse systeem van budgethoudende huisartsen (GP Fundhol-

ding) bleek succesvol, bijvoorbeeld bij het beperken van de kosten voor onder meer geneesmiddelen, het terugbrengen van de wachttijden voor specialistische hulp, het verlagen van het aantal ziekenhuisverwijzingen en het verhogen van het aantal dagbehandelingen. Blue Plus vertoonde een kostendaling, zowel op het niveau van de verzekeraar als op het niveau van de arts, hoewel het effect verschilde per manier van betalen. Minder succesvol was United Healthcare, hoewel de kosten per verzekerde spectaculair afnamen nadat structurele organisatorische en financiële wijzigingen waren doorgevoerd. Of deze daling het gevolg was van veranderd gedrag van de artsen of dat deze het gevolg was van bijvoorbeeld een selectieve uitstroom van verzekerden met hoge kosten, is onduidelijk.

Hoewel de bevindingen wijzen op een onderscheidend effect van systemen met financiële-risicodeling op het gedrag van huisartsen als agenten voor hun patiënten, worden definitieve conclusies vooral bemoeilijkt door methodologische problemen. Het ontbreekt vaak aan een gedegen onderzoeksopzet. Dit maakt het moeilijk om te beoordelen in hoeverre het gedrag van de arts wordt beïnvloed door andere (financiële) prikkels, door andere managed-caretechnieken of door externe factoren. Verder leveren studies naar systemen met financiële-risicodeling vaak weinig informatie over de kwaliteit van zorg, terwijl dit een belangrijke uitkomstmaat is voor het beoordelen van het functioneren van de artsen als agenten voor hun patiënten.

8 Naar een systeem met financiële-risicodeling

Er kunnen belangrijke lessen worden getrokken uit de analyse van de theoretische en praktijkmodellen van systemen met financiële-risicodeling. Om verschillende redenen zijn deze lessen echter niet beslissend genoeg om een normatief model voor een systeem met financiële-risicodeling op te stellen. Ten eerste is het aantal geanalyseerde praktijkvoorbeelden beperkt. De uit deze voorbeelden getrokken conclusies kunnen daarom niet eenvoudigweg worden toegepast op andere gezondheidszorgsystemen. Verder is econometrisch onderzoek cruciaal voor het ontwerpen van een model van financiële-risicodeling. De verscheidene aspecten van financiële-risicodeling bepalen gezamenlijk de hoeveelheid risico die van de zorgverzekeraar wordt overgeheveld naar de huisarts en dienen nauwkeurig in evenwicht te worden gebracht. Ten slotte zal de uiteindelijke vorm van financiële-risicodeling bijvoorbeeld per land en per gezondheidszorgsysteem verschillen, afhangen van de rol en positie van de huisartsen, afhangen van de preferenties van de huisartsen en worden bepaald door de doelen die de zorgverzekeraar nastreeft. Toch hebben we een onderscheid gemaakt tussen het ontwerp voor een eerste regeling met financiële-risicodeling en de uiteindelijke regeling. Een eerste regeling wordt gekenmerkt door:

- een beperkt risicopakket (bijvoorbeeld alleen de kosten van geneesmiddelen);
- een risicocontract voor minimaal vijftig procent van de praktijkpopulatie;
- een norm die (gedeeltelijk) is gebaseerd op historische gegevens, om een relatief vloeiende overgang naar een systeem van financiële-risicodeling mogelijk te maken;
- een eenvoudig bonussysteem;
- een ‘excess of loss per risk’ in combinatie met een ‘stop loss’ contract en een groepspraktijk of waarneemgroep als risicopool.

Een minder behoudende regeling kan worden overwogen indien een vloeiende overgang naar een systeem van risicodeling is gemaakt, zowel zorgverzekeraars als artsen ervaring hebben opgedaan met risicodeling en de effecten op gedrag en uitkomsten bekend zijn. De regeling voor risicodeling kan dan worden gekenmerkt door:

- een breed risicopakket waarvan dure behandelingen en behandelingen met een open einde zijn uitgezonderd;
- een risicocontract voor bij voorkeur de gehele praktijkpopulatie;
- een norm die is gebaseerd op de behoefte van patiënten aan zorg, die een zekere voorspellende waarde heeft voor toekomstige kosten van gezondheidszorg, die is gebaseerd op een optimaal niveau van zorg en die fungeert als doel;
- een meer geavanceerd bonussysteem met een variabele bonus, bij voorkeur een omgekeerd evenredig bonussysteem in combinatie met een doelstellende norm;
- een 'excess of loss per risk' in combinatie met een 'stop loss' contract en een groepspraktijk of waarneemgroep als risicopool.

Wat betreft het bonussysteem zal een zorgverzekeraar waarschijnlijk een voorkeur hebben voor een omgekeerd evenredige bonus met een norm als doel, omdat deze de te bereiken doelen expliciet maakt en ongewenst gedrag van de arts minder waarschijnlijk maakt. Bij een (omgekeerd) evenredig systeem moet de bonus of de malus substantieel zijn, maar minder dan honderd procent. Circa vijftig procent lijkt dan een goed uitgangspunt, maar uiteindelijk zal een besluit over het percentage in samenhang met de overige aspecten van financiële-risicodeling moeten worden genomen.

Het is van groot belang om alle aspecten van een regeling met financiële-risicodeling, zowel de vijf hoofdaspecten als de overige aspecten, zeer zorgvuldig op elkaar af te stemmen. Bovendien zal financiële-risicodeling vergezeld moeten gaan van andere technieken uit de managed-carecyclus. Idealiter komen alle drie de fases uit de managed-carecyclus voor in het ontwerp van de regeling. Slechts dan kan een regeling voor risicodeling worden verkregen die de zorgverzekeraar helpt om een doelmatige en kwalitatief hoogwaardige zorg te bereiken.

9 Epiloog

Hoewel het gedrag van huisartsen mede bepalend is voor de doelmatigheid van de gezondheidszorg, is er betrekkelijk weinig empirisch en theoretisch conceptueel onderzoek voorhanden naar de (effecten van) financiële prikkels die dit gedrag kunnen beïnvloeden. Dit geldt zeker voor financiële-risicodeling tussen zorgverzekeraars en huisartsen, maar ook voor de bredere set managed-caretechnieken die door een zorgverzekeraar kan worden gebruikt om te trachten het gedrag van huisartsen te beïnvloeden. Door een conceptueel raamwerk voor financiële-risicodeling op te stellen en door dit vervolgens te gebruiken voor de analyse van praktijkvoorbeelden en het trekken van lessen hieruit, hebben we een bijdrage willen leveren aan de literatuur over honoreringssystemen voor huisartsen. Door een verbinding te leggen tussen agency-theorie en managed care en door de verschillende technieken te classificeren en met elkaar in relatie te brengen in de managed-care cyclus, hebben we tevens een bijdrage willen leveren aan de literatuur over managed care. Deze bijdragen zijn relevant omdat honoreringssystemen, financiële-

risicodeling en de overige managed-caretechnieken belangrijke vraagstukken zijn in de wereldwijde discussie over verbetering van de doelmatigheid van de zorg.

In hoofdstuk 7 hebben we enkele voorbeelden van systemen met financiële-risicodeling in Nederland beschreven. Hoewel er in het verleden wel meerdere projecten zijn opgestart en invoering van dergelijke systemen in Nederland al vaak is voorgesteld (zie paragraaf 7.2.1), is er in Nederland niet veel ervaring opgedaan met financiële-risicodeling tussen zorgverzekeraars en huisartsen. In hoofdstuk 6 hebben we beargumenteerd dat er goede redenen zijn om het verzekeringsrisico en het risico van imperfect agentschap tussen beide partijen te delen. De per 1 januari 2006 ingevoerde Zorgverzekeringswet en de nieuwe Wet Marktordening Gezondheidszorg geven zorgverzekeraars en huisartsen daartoe ook meer ruimte en mogelijkheden. Daarnaast spelen er andere ontwikkelingen die ertoe bijdragen dat voor beide partijen de tijd rijp lijkt te zijn om in overleg te treden over een systeem waarin de risico's worden gedeeld. Voorbeelden hiervan zijn de professionalisering van de onderhandelingen tussen beide partijen, de overgang naar grotere samenwerkingsverbanden van huisartsen en de belangstelling onder in ieder geval een deel van de huisartsen voor het ondernemerschap.

De tijd lijkt echter niet rijp te zijn om tot grootschalige invoering van financiële-risicodeling voor vervolgkosten over te gaan. Het verdient aanbeveling om op bescheiden schaal te starten in de vorm van experimenten. Hiervoor zijn in ieder geval twee belangrijke redenen te noemen. Een eerste reden is dat systemen van financiële-risicodeling veel aspecten kennen waarop kan worden gevarieerd. Daarbij kunnen de vele variaties verschillend uitpakken voor de mate waarin het risico wordt overgedragen aan de huisarts en voor de wijze waarop de huisarts op de prikkels reageert. In hoofdstuk 6 hebben we vier mogelijke effecten onderscheiden. Enerzijds kan een positief effect op de doelmatigheid van de zorg optreden. Anderzijds kan de huisarts ook ongewenst gedrag vertonen door gunstige risico's te selecteren, de kosten door te schuiven en op de kwaliteit van zorg te bezuinigen. Belangrijk doel van de experimenten is dan ook het verkrijgen van meer inzicht in de mate waarin de verschillende organisatorische en financiële structuren tot dergelijke effecten leiden.

Een tweede reden om niet tot grootschalige invoering over te gaan maar om te starten met experimenten is dat, hoewel van ieder systeem van honoreren financiële prikkels uitgaan, een discussie over financiële prikkels gevoelig ligt in de beroepsgroep. Ondanks deze gevoeligheid zijn er verschillende redenen om te veronderstellen dat invoering van financiële-risicodeling in Nederland haalbaar is:

- Het volledige abonnementssysteem in de voormalige Ziekenfondssector is een voorbeeld van een systeem waarin decennia lang financiële-risicodeling heeft plaatsgevonden, al had het abonnement geen betrekking op de kosten van vervolgzorg.
- Nederlandse artsen hebben in het verleden actief met dergelijke systemen gewerkt, al dan niet in een experiment (bijvoorbeeld 'Zaanland', 'Amsterdam' en 'Tilburg'; zie hoofdstuk 7).
- Vanuit de beroepsgroep zelf zijn voorstellen voor financiële-risicodeling ontwikkeld (bijvoorbeeld door de Paritaire Werkgroep Huisartsenzorg; zie paragraaf 7.2.1).
- Huisartsen in andere landen werken vaak op vrijwillige basis met dergelijke systemen dan wel hebben ermee gewerkt.

- De huidige belangstelling onder huisartsen om met zorgverzekeraars afspraken te maken over eerstelijns DBC's voor ketenzorg.

Ongeacht deze positieve voortekenen is het voor goed functionerende systemen belangrijk om ze in nauwe samenspraak met de beroepsgroep vorm te geven en aan te sluiten bij de professionele normen en waarden van de beroepsgroep. Door op kleine schaal en met een groep enthousiaste huisartsen te starten met experimenteren, kan eerst worden onderzocht wat de effecten in de praktijk zijn. Bij positieve effecten is bredere invoering dan wellicht acceptabel voor een grotere groep huisartsen. Andere belangrijke doelen van de experimenten zijn dus het in een vroeg stadium van het ontwerp betrekken van de beroepsgroep en het vergroten van het draagvlak voor invoering.

Centraal in de experimenten staan de verschillende mogelijkheden voor financiële-risicodeling voor bepaalde kosten van vervolgzorg. In het voorgaande hebben we echter gesteld dat risicodeling één van de mogelijkheden is om het gedrag van huisartsen te sturen en dat deze mogelijkheid vergezeld zal moeten gaan van andere technieken uit de managed-carecyclus. Idealiter worden dan ook bij de opzet van het experiment de drie opeenvolgende fases uit de cyclus, te weten selecteren, sturen en monitoren, in samenhang vormgegeven. Op deze onderdelen wordt in het onderstaande ingegaan. Daarnaast gelden voor de opzet van experimenten andere aandachtspunten, waaronder de samenwerking tussen verschillende beroepsgroepen, doelmatigheid versus bureaucratisering en administratieve lasten, de gegevensverzameling, automatisering en privacy (zie ook Breedveld et al. 1994). Deze blijven hier buiten beschouwing.

Op basis van de in eerdere hoofdstukken beschreven theoretische en empirische bevindingen en gelet op de huisartsenzorg in Nederland, komen meerdere modellen in aanmerking voor een experiment. Hier worden twee modellen voorgesteld:

- Een bonussysteem voor solopraktijken.
- Een bonussysteem voor groepspraktijken.

De bonussystemen in de experimenten hebben betrekking op bepaalde kosten van vervolgzorg en worden naast de basishonorering van de huisarts toegepast. Het zou interessant zijn om in een apart experiment te onderzoeken welke varianten van basishonorering (zoals een abonnement, betaling per consult of gemengd systeem) gecombineerd met een bepaalde vorm van risicodeling voor vervolgkosten tot welke resultaten leiden. Ervan uitgaande dat een huisarts zich door een abonnementssysteem eerder geprikkeld voelt tot doorverwijzen dan door een consultsysteem, zou bij een volledig abonnement de bonus vooral moeten zijn gericht op het terugdringen van het aantal verwijzingen. Een andere hypothese zou kunnen zijn dat een vergelijkbaar bonussysteem in combinatie met een volledig consultsysteem, leidt tot een te hoog aantal onterecht niet verwezen patiënten. Bij de opzet van de experimenten is er echter van uitgegaan dat ze worden toegepast in aanvulling op de huidige gemengde honoreringsystematiek van huisartsen, bestaande uit inschrijf-, consult- en moduletarieven.

Experimenteren met een bonussysteem voor solopraktijken

Hoewel het aantal solopraktijken al jaren terugloopt, is het in Nederland de meest voorkomende praktijkvorm. Dat maakt de solopraktijk nog steeds tot een relevante praktijk-

vorm voor een experiment. Het bonus-malus experiment in Tilburg heeft bovendien laten zien dat een experiment met vooral solopraktijken succesvol kan zijn. Ook de ervaringen van United Healthcare zijn, hoewel minder succesvol, in dit verband relevant.

Selecteren
Het eerste onderdeel betreft de selectie van individuele artsen (en dus van praktijken) die aan het experiment willen deelnemen. Drie selectiemethoden liggen daarbij voor de hand. Een eerste methode is zelfselectie. Door vooraf goed te specificeren wat het doel is van het experiment en hoe het wordt opgezet, zullen naar verwachting artsen geïnteresseerd zijn die doel en opzet onderschrijven. Een tweede methode betreft het analyseren van declaratiegegevens. De overgang naar een gedeeltelijk consultsysteem in de huisartsenzorg in combinatie met de toename van het elektronische declaratieverkeer maakt een eerste analyse van het gedrag van zorgaanbieders betrekkelijk eenvoudig. Voor een diepere analyse zijn meer gegevens nodig. Daarbij gaat het vooral om patiëntkenmerken, morbiditeit, artsspecifieke kenmerken, omgevingsfactoren en (de methode van verzameling van) de gebruikte data (NIVEL/RIVM 2004). Een derde methode is selectie op basis van de kennis en ervaringen van accountmanagers huisartsenzorg werkzaam bij de betreffende zorgverzekeraar. Deze hebben veelal enig zicht op de praktijkvoering door huisartsen in hun regionale werkgebied.

Sturen
Poortwachterschap is een voor de hand liggende combinatie met een financiële prikkel voor de kosten van vervolgzorg. Voor de Nederlandse huisarts is dit al bij wet geregeld. Andere sturingstechnieken die hier voor eerste experimenten worden voorgesteld hebben vooral een informerende en overtuigende werking, zoals het opstellen van praktijkprofielen en het stimuleren van het volgen van NHG-standaarden. Het gebruiken van de derde groep sturingstechnieken, namelijk sturing door regels en autoriteit wordt hier vooralsnog afgeraden. Deze lijken in de Nederlandse verhoudingen niet opportuun en passen niet bij de professionele, onafhankelijke status van de Nederlandse huisarts.

Monitoren
Het monitoren is om twee redenen een onmisbaar onderdeel van de experimenten. Ten eerste verkrijgt de zorgverzekeraar hiermee de noodzakelijke informatie over het gedrag van de deelnemende huisartsen of de uitkomst van de processen waaraan ze hebben bijgedragen. Ten tweede levert het monitoren feedbackinformatie op voor deelnemende huisartsen waardoor ze kunnen worden geïnformeerd over hun (relatieve) prestaties. Om goed te kunnen monitoren is een nauwkeurige, uniforme registratie van verrichtingen en declaraties essentieel. Dit vergt een uniform (gebruik van een) registratiesysteem en voorlichting aan deelnemende huisartsen. Het probleem van onderregistratie zal sinds de invoering van het gemengde honoreringssysteem (per 1 januari 2006) zijn verminderd voor declarabele verrichtingen, maar blijft een gevaar voor niet-declarabele verrichtingen.

Betrouwbare monitoring vergt verschillende betrouwbare metingen. Dit verkleint de kans op gewenst gedrag op de punten waarop wordt gemeten. Daarnaast is uit onderzoek

gebleken dat de prestaties van een arts bij de behandeling van een bepaalde ziekte geen goede indicatie vormen voor de prestaties op andere delen van het vakgebied.

Belangrijk onderdeel van het monitoren betreft het meten van de effecten van het systeem van risicodeling op de kwaliteit van zorg. Er zullen tussen zorgverzekeraar en huisartsen(groep) nadere afspraken moeten worden gemaakt over de te hanteren indicatoren, zoals bij- en nascholing, openingstijden en bereikbaarheid, praktijkvoorzieningen, het volgen van standaarden en de gezondheidstoestand en tevredenheid van patiënten.

Belangrijk aandachtspunt betreft de optimale combinatie van de verschillende managed-caretechnieken. Met andere woorden, welke technieken zijn complementair en welke technieken zijn substitueerbaar? Het betreft dan vooral de vraag welke sturingstechnieken dienen te worden gekozen. We hebben immers beargumenteerd dat de drie fases uit de managed-carecyclus allemaal aan bod zouden moeten komen. Bij de selectie- en de monitoringsfase zijn er, in tegenstelling tot bij de sturingsfase, echter niet veel opties. Wel zou de keuze voor zelfselectie of voor selectie door de zorgverzekeraar een verschil kunnen maken. Bij zelfselectie betreft het naar verwachting vooral gemotiveerde huisartsen die zich kunnen vinden in de gestelde doelen van een systeem van risicodeling. De hypothese zou dan kunnen zijn dat bij zelfselectie sturende technieken (zoals allerlei vormen van protocollen, machtigingen of verplichte second opinions) minder nodig zijn dan bij selectie door de zorgverzekeraar. Dit hangt echter weer af van de opzet van het systeem van risicodeling. Een systeem zou ook tot antiselectie kunnen leiden: juist ondoelmatig werkende huisartsen worden aangetrokken, omdat er voor hen meer winst valt te behalen dan voor doelmatig werkende collega's. Eerder is al gewezen op de logische combinatie van financiële prikkels voor vervolgzorg en het poortwachterschap. De combinatie van financiële prikkels en sterk sturende technieken ligt minder voor de hand. Bij financiële prikkels worden bepaalde keuzes voor de arts relatief aantrekkelijker gemaakt, terwijl bij bijvoorbeeld protocollen of machtigingen de keuzevrijheid juist sterk wordt ingeperkt of zelfs afwezig is. Daarnaast is er al op gewezen dat sterk sturende technieken in de Nederlandse verhoudingen niet opportuun lijken en niet lijken te passen bij de professionele, onafhankelijke status van de Nederlandse huisarts.

Financiële-risicodeling

Bij het bonus-malus experiment in Tilburg en bij United Healthcare was sprake van een breed risicopakket. Belangrijk verschil was de mate waarin de huisartsen risico liepen. Dit verschil was vooral het gevolg van het feit dat in 'Tilburg' huisartsen voor tweederde van de praktijkpopulatie risico liepen, terwijl het bij United Healthcare slechts om zeer kleine aantallen patiënten ging. Bij de huidige Nederlandse marktverhoudingen zal al snel 50 procent van de praktijkpopulatie zijn verzekerd bij de regionale zorgverzekeraar. De omvang van de relevante praktijkpopulatie zou verder kunnen worden vergroot door een andere zorgverzekeraar bij het experiment te betrekken. Ook zou een deel van de overige patiënten in het experiment kunnen worden betrokken door de contracten met en de betalingen van de desbetreffende zorgverzekeraars via een tussenliggende organisatie te laten verlopen.

In de praktijk zal het niet wenselijk zijn om solopraktijken te belasten met de financiële, organisatorische en personele lasten van een meer gecompliceerd systeem van financiële-risicodeling. Daarom is er juist bij deze variant voor gekozen het risicopakket in het experiment te beperken tot slecht één enkel onderdeel, namelijk geneesmiddelen. De keuze juist voor geneesmiddelen sluit ook aan bij de vele rapporten die er over het geneesmiddelenbeleid zijn geschreven en waarin voorstellen zijn gedaan om financiële prikkels in te voeren gericht op een doelmatig voorschrijfbeleid door huisartsen (zie hiervoor paragraaf 7.2.1). Voordeel is bovendien dat in het experiment kan worden gevolgd of het beperken tot geneesmiddelen bijvoorbeeld leidt tot een stijging van het aantal verwijzingen naar het ziekenhuis. Verder maakt een keuze voor alleen geneesmiddelen het systeem eenvoudig en overzichtelijk.

Dat het meest complexe onderdeel van het experiment het vaststellen van een norm is op basis waarvan achteraf de bonus (en eventueel de malus) wordt vastgesteld, is wel gebleken uit de in hoofdstuk 7 beschreven ervaringen. In hoofdstuk 6 zijn verschillende manieren beschreven waarop middels kostengegevens tot een norm zou kunnen worden gekomen. Een norm op basis van historische kosten of op basis van gemiddelde kosten is de meest simpele methode om tot een norm te komen. Historische kosten hebben bovendien als voordeel dat invoering van risicodeling niet tot grote schommelingen leidt. Vanwege de nadelen van deze methoden, zoals een oneerlijke verdeling van middelen over de huisartsen en het gevaar van risicoselectie, verdient het sterk de voorkeur om de norm te verfijnen op basis van verdeelkenmerken. Onderzocht zou moeten worden in hoeverre de verdeelkenmerken zoals die nu worden gebruikt voor de bepaling van normuitkeringen voor zorgverzekeraars, ook kunnen worden gebruikt voor het vaststellen van de norm voor huisartsen. Het gaat hierbij om verdeelkenmerken als leeftijd, geslacht, woonplaats, aard van het inkomen, gebruik van bepaalde geneesmiddelen (via Farmaceutische Kosten Groepen of FKG's) en specifieke aandoeningen (via Diagnose Kosten Groepen of DKG's). Verdeelkenmerken die dan mogelijk in aanmerking komen voor de norm voor huisartsen zullen dan kenmerken dienen te zijn waarop huisartsen geen of moeilijk invloed kunnen uitoefenen. Deze eis kan ertoe leiden dat een verdeelkenmerk dat wel wordt gebruikt in het normuitkeringensysteem niet of op een andere manier wordt gebruikt voor het vaststellen van de norm voor huisartsen. Ook het omgekeerde is mogelijk.

Uit de in hoofdstuk 7 beschreven ervaringen is gebleken dat bepaalde geneesmiddelen van het risicopakket worden uitgesloten. Het gaat dan om zeer kostbare geneesmiddelen of geneesmiddelen voor patiënten met hoge zorgkosten. Interessante vraag voor een experiment is of alleen de kosten van geneesmiddelen die door de huisarts zijn voorgeschreven bepalend zijn, of dat ook de kosten van door specialisten geïnitieerde behandelingen dienen te worden meegenomen. Zie voor de overwegingen verder de discussie in paragraaf 7.2.3.

Het grote voordeel van een systeem van financiële-risicodeling voor huisartsen is dat zij delen in de besparingen door doelmatig voorschrijfgedrag; besparingen die anders volledig bij de zorgverzekeraar terechtkomen. Hoewel arbitrair, lijkt een bonus van maximaal 50% van de bespaarde kosten in beginsel een goed uitgangspunt. Huisartsen worden geprikkeld doordat ze hiermee substantieel delen in de eventuele besparingen. De prikkels zijn echter niet gemaximeerd teneinde ongewenst gedrag tegen te gaan. In de

experimenten is dit echter een percentage waarin zou kunnen worden gevarieerd.

Een andere vraag die in de experimenten aan de orde moet komen is of er een malus wordt gehanteerd. Verschillende overwegingen spelen hierbij een rol. Enerzijds vermindert een malus de aantrekkelijkheid van deelname aan een systeem van risicodeling. Anderzijds is er door de combinatie van bonus en malus daadwerkelijk sprake van deling van de risico's. Hier wordt voorgesteld om in de experimenten met een fictieve malus te werken. De eventuele malus wordt wel berekend en aan de desbetreffende artsen bekendgemaakt, maar leidt vooralsnog niet tot een daadwerkelijke korting of terugvordering. Uit de totale bevindingen van de experimenten moet vervolgens blijken waardoor de eventuele malussen zijn ontstaan, zoals door ondoelmatig gedrag van de desbetreffende artsen, door foutief vastgestelde normen of door toeval.

Het risico kan verder worden beperkt door een 'excess of loss per risk'. Hierdoor wordt voorkomen dat enkele patiënten met hoge kosten voor geneesmiddelen de hoogte van de bonus sterk negatief beïnvloeden. Aangezien is voorgesteld om de malus niet daadwerkelijk te verrekenen, zijn andere maatregelen, zoals een 'stop loss', niet nodig. De huisartsen maken ook geen deel uit van een risicopool.

De bonus wordt in dit model beschouwd als een vorm van extra inkomsten, waarbij het aan de huisarts is om te bepalen of deze als inkomen worden beschouwd of (gedeeltelijk) worden geïnvesteerd in de zorgverlening.

Tabel 1. Bonussysteem solopraktijk

Huisartsen	Zelfstandig gevestigde huisartsen werkzaam in een solopraktijk.
Risicopakket	Door huisarts voorgeschreven geneesmiddelen. Nader vast te stellen geneesmiddelen of patiënten met te verwachten hoge kosten uitsluiten. Aandachtspunt: kosten van door specialisten geïnitieerde farmaceutische behandelingen.
Omvang praktijkpopulatie	Circa 50 procent van praktijkomvang (uitgaande van een normpraktijk circa 1.200 patiënten), eventueel te verhogen door (al dan niet via tussenliggende organisatie) verzekerden van andere zorgverzekeraars te betrekken.
Norm	Idealiter een norm vastgesteld op basis van verdeelkenmerken zoals leeftijd, geslacht, woonplaats, aard van het inkomen, FKG's en/of DKG's. Indien dit leidt tot grote verschuivingen, eventueel in de overgangsperiode gecombineerd met historische kosten.
Bonussysteem	Bonus van maximaal 50 procent van bespaarde kosten. Fictieve malus.
Risico beperken	'Excess of loss per risk'. Geen risicopool.
Besteding bonus	Beslissing huisarts: beschouwen als inkomen en/of investeren in de zorgverlening.

Experimenteren met een bonussysteem voor groepspraktijken

Een toenemend aantal huisartsen is werkzaam in een mono- of multidisciplinaire groepspraktijk. Dit biedt de mogelijkheid om een experiment op te zetten waarbij de zorgverze-

keraar het risico niet deelt met de individuele huisartsen maar met de groepspraktijk. We laten hierbij in het midden of sprake is van een mono- of een multidisciplinaire groepspraktijk en zetten het experiment alleen op voor de in de groepspraktijk werkzame huisartsen. In het eerste experiment is de individuele huisarts contractspartij en is er geen sprake van een tussenliggende organisatie. In het tweede experiment contracteert de zorgverzekeraar de groepspraktijk. De groepspraktijk vormt daarmee een tussenliggende organisatie tussen de zorgverzekeraar en de individuele huisarts. Hoewel afhankelijk van eventuele afspraken daarover tussen zorgverzekeraar en tussenliggende organisatie, is het bij dergelijke modellen in principe de laatste die bepaalt hoe de relatie tussen de tussenliggende organisatie en individuele zorgaanbieder wordt vormgegeven.

Voor het experiment is het essentieel om vast te stellen wie de relatie tussen de tussenliggende organisatie en de individuele huisarts vormgeeft. Er zijn dan twee opties. De eerste optie is dat in het experiment alleen de relatie tussen zorgverzekeraar en groepspraktijk wordt vastgelegd. Het is dan aan de groepspraktijk om te bepalen hoe de afspraken met de zorgverzekeraar worden vertaald in afspraken met de individuele huisartsen. In het experiment kan dan worden onderzocht hoe een groepspraktijk reageert op de met de zorgverzekeraar gemaakte afspraken en vervolgens wat het effect is van de afspraken tussen groepspraktijk en huisartsen op de zorg. Deze optie is vooral geschikt als de zorgverzekeraar slechts is geïnteresseerd in het overdragen van een deel van het risico. In het in hoofdstuk 4 opgestelde raamwerk gaan we er echter vanuit dat de zorgverzekeraar als agent van de verzekerde/patiënt optreedt en daarom invloed wil kunnen uitoefenen op de arts-patiëntrelatie. We hebben tevens gesteld dat deze invloed vooral via de arts zal plaatsvinden. Daarom wordt hier een andere optie voorgesteld, namelijk dat in het experiment tevens wordt vastgelegd hoe de relatie tussen huisartsengroep en individuele huisartsen wordt vormgegeven.

Selecteren

Aangezien de groepspraktijk contractspartij is, gaat het bij dit experiment in de eerste plaats om de selectie van praktijken (en dus niet van individuele artsen) die aan het experiment willen deelnemen. Ook hier geldt dat zelfselectie, het analyseren van declaratiegegevens en selectie op basis van de kennis en ervaringen van accountmanagers huisartsenzorg voor de hand liggen. Aangezien ook in groepspraktijken huisartsen vaak individueel declareren, zal de analyse van declaratiegegevens toch informatie over de individuele artsen opleveren. De selectie op groepsniveau kan dan toch worden beïnvloed door informatie op individueel niveau.

Bij selectie door de zorgverzekeraar op groepsniveau zou de selectie van huisartsen op individueel niveau (de toetreding tot de groepspraktijk) aan de groep kunnen worden overgelaten. Gezien de agentschapsfunctie van de zorgverzekeraar ligt het echter wel voor de hand dat deze in het contract met de groepspraktijk nadere (kwaliteits)eisen aan de in de praktijk werkzame huisartsen stelt. De daadwerkelijke selectie vindt dan wel door de groep plaats.

Sturen
De sturingstechnieken komen deels overeen met de technieken in het eerste experiment: poortwachterschap in combinatie met financiële prikkels gericht op de kosten van vervolgzorg, het opstellen van praktijkprofielen en het stimuleren van het volgen van NHG-standaarden. Het gebruik door de zorgverzekeraar van de derde groep sturingstechnieken, namelijk sturing door regels en autoriteit, ligt ook hier om de eerder genoemde redenen niet voor de hand.

In de relatie tussen groep en individuele huisartsen kunnen afwijkende technieken worden gehanteerd. Zo is een punt van aandacht welk systeem voor de basishonorering wordt gebruikt. Zoals gesteld gaan we uit van de huidige gemengde honoreringsystematiek van huisartsen, bestaande uit inschrijf-, consult- en moduletarieven. In sommige groepspraktijken werken echter ook huisartsen in loondienst. Mogelijk leidt een salaris in combinatie met een bepaalde vorm van risicodeling tot een ander effect dan de gemengde honorering in combinatie met dezelfde vorm van risicodeling. Daarnaast kunnen de nabijheid van collega's en druk om bepaalde groepsnormen, nadere afspraken, protocollen of standaarden te respecteren van invloed zijn op het gedrag van de artsen. Bij dienstverband kan bovendien een gezagsrelatie aan de orde zijn.

Monitoren
Het monitoren kan zowel op groepsniveau als op individueel niveau plaatsvinden. Als huisartsen bijvoorbeeld individueel declareren, levert de analyse van declaratiegegevens informatie op individueel niveau. De zorgverzekeraar kan ook de contractpartij (zijnde de groepspraktijk) om nadere gegevens op individueel niveau vragen. Het aanleveren van deze gegevens kan wel degelijk in het belang van de artsen zijn, bijvoorbeeld indien deze gegevens nodig zijn om de bonussen te berekenen of indien ze informatie kunnen opleveren over de relatieve prestaties van de artsen.

Verder geldt ook hier dat er tussen zorgverzekeraar en huisartsen(groep) nadere afspraken moeten worden gemaakt over de te hanteren indicatoren voor de kwaliteit van zorg.

Financiële-risicodeling
Omdat de opzet van het onderdeel financiële-risicodeling voor een deel overeenkomt met de opzet in het experiment voor solopraktijken, beperken we ons hier tot de verschillen. Een eerste verschil is dat in dit experiment de deelnemende huisartsen voor een breed risicopakket verantwoordelijk zijn, bestaande uit verwijzingen naar de fysiotherapeut, verwijzingen naar de polikliniek en de door de huisarts voorgeschreven geneesmiddelen. Nader vast te stellen zeer kostbare geneesmiddelen of geneesmiddelen voor patiënten met hoge zorgkosten worden uitgesloten, evenals spoedeisende hulp. Ook hier geldt als aandachtspunt het al dan niet verantwoordelijk maken van de huisartsen voor de kosten van door specialisten geïnitieerde behandelingen. Belangrijke voordelen van een breed pakket zijn dat het risico over een breder deel van de zorg kan worden gedeeld, dat het integrale zorg en substitutie bevordert en dat het moeilijker wordt om zorg waarvoor de arts (deels) financiële verantwoordelijkheid draagt te vervangen door zorg waarvoor hij geen verantwoordelijkheid draagt. Doordat het om groepspraktijken met verantwoordelijkheid voor

een breed risicopakket gaat, zou dit experiment op termijn kunnen uitgroeien tot een model van budgethouderschap (vergelijkbaar met GP Fundholding). Het experiment sluit bovendien aan bij het groeiende aantal groepspraktijken en bij de toekomstvisie van de beroepsgroep op huisartsenzorg. Ook sluit het aan bij ontwikkelingen in de eerstelijnszorg waarbij partijen integrale afspraken over een deel van de zorg willen maken (bijvoorbeeld afspraken over ketenzorg).

Tabel 2. Bonussysteem groepspraktijk

Huisartsen	Zelfstandig gevestigde huisartsen werkzaam in een groepspraktijk.
Risicopakket	Breed samengesteld, bestaande uit verwijzingen fysiotherapie, verwijzingen polikliniek, door huisarts voorgeschreven geneesmiddelen. Nader vast te stellen geneesmiddelen of patiënten met te verwachten hoge kosten uitsluiten, evenals spoedeisende hulp. Aandachtspunt: kosten van door specialisten geïnitieerde farmaceutische behandelingen.
Omvang praktijkpopulatie	Circa 50 procent van de gezamenlijke praktijkomvang (uitgaande van een groepspraktijk met 10.000 patiënten circa 5.000 patiënten), eventueel te verhogen door (al dan niet via tussenliggende organisatie) verzekerden van andere zorgverzekeraars te betrekken.
Norm	Idealiter een norm vastgesteld op basis van verdeelkenmerken zoals leeftijd, geslacht, woonplaats, aard van het inkomen, FKG's en/of DKG's. Indien dit leidt tot grote verschuivingen, eventueel in de overgangsperiode gecombineerd met historische kosten.
Bonussysteem	Bonus van maximaal 50 procent van bespaarde kosten. In eerste instantie fictieve malus. In tweede instantie een malus van maximaal 50 procent.
Risico beperken	'Excess of loss per risk', later in combinatie met een 'stop loss'. Een risicopool bestaande uit de drie tot vijf huisartsen in de groepspraktijk.
Besteding bonus	Investeren in de (kwaliteit van de) zorgverlening.

Ook hier wordt in eerste instantie gewerkt met een fictieve malus. Nadat de aanloopproblemen met het vaststellen van een norm zijn opgelost, wordt echter zo snel mogelijk overgestapt op een gecombineerde bonus/malus van maximaal 50 procent. De gecombineerde bonus/malus is tevens de opstap naar een toekomstig systeem van budgethouderschap. Daarbij bedragen bonus en malus 100 procent tot een nader vast te stellen maximum (maximale winst en 'stop loss').

Vanwege de grotere financiële risico's maken de huisartsen uit de groepspraktijk deel uit van een risicopool. De risico's worden hiermee gespreid over de groep deelnemende huisartsen. Daarbij zou ervoor kunnen worden gekozen om de bonus of de (fictieve) malus op individueel niveau te berekenen, maar om de risico's zoveel mogelijk te spreiden verdient berekening op groepsniveau de voorkeur. Een andere reden om bonus en malus op groepsniveau vast te stellen is dat in dit experiment ervoor is gekozen de besteding van de bonus te beperken tot investeringen in de (kwaliteit van de) zorgverlening. Deze opzet is vergelijkbaar met die van GP Fundholding. De inkomensprikkel die uitgaat

van verantwoordelijkheid voor een breed risicopakket wordt hierdoor gedempt, hoewel ook investeringen in de zorgverlening indirect tot inkomensverbeteringen kunnen leiden (bijvoorbeeld doordat de praktijk aantrekkelijker wordt voor patiënten of via een productiestijging door investeringen in extra personeel). In een groepspraktijk kunnen investeringen in de zorgverlening makkelijker gezamenlijk plaatsvinden dan in een samenwerkingsverband van solopraktijken (bijvoorbeeld door personeel, ruimte, inventaris en instrumentarium te delen). Daarnaast zullen discussies in de groep huisartsen over bijvoorbeeld werkbelasting, praktijkstijl en inkomensverschillen worden beperkt. Enerzijds is het wenselijk dat financiële-risicodeling leidt tot dergelijke discussies, tot meer uniformering en reductie van praktijkvariatie. Anderzijds dienen ingewikkelde discussies over bijvoorbeeld de relatieve bijdragen van de huisartsen aan de besparingen niet verstorend te werken op de onderlinge verhoudingen.

Op grond van de hier beschreven theoretische en empirische bevindingen menen we dat er goede redenen zijn om financiële-risicodeling tussen zorgverzekeraars en huisartsen ook in Nederland in te voeren. Nader econometrisch onderzoek en de bevindingen uit experimenten zullen moeten uitwijzen of invoering van financiële-risicodeling ook daadwerkelijk haalbaar en wenselijk is.

DANKWOORD

‘Soms zit het mee, soms zit het tegen.’ Dat de vroegere reclameslogan van een Nederlandse verzekeraar ook van toepassing kan zijn op het schrijven van een proefschrift over de relatie tussen een zorgverzekeraar en een huisarts, heb ik mogen ervaren. Soms zat het persoonlijk mee en zat het proefschrift ook mee. Het heeft echter door persoonlijke omstandigheden een paar keer flink tegen gezeten, waardoor het bij het schrijven niet echt mee zat. Ook kwam het voor dat het persoonlijk erg mee zat (zoals bij het krijgen van mijn twee lieve kindjes), waardoor het bij het schrijven van het proefschrift juist weer tegen zat. Er was voor anderen waarschijnlijk geen peil op te trekken.

Wynand van de Ven en Erik Schut bedank ik voor de geweldige begeleiding. Zoals dat hoort op het multidisciplinaire instituut Beleid en Management Gezondheidszorg, was de aanpak multidisciplinair. Economie, Engels, methodologie en psychologie zijn voorbeelden van disciplines die bij de begeleiding aan de orde kwamen. Beiden wisten altijd de tijd te vinden om in alle rust mee te denken, om mij te stimuleren en om mijn stukken tekst snel van waardevol commentaar te voorzien. Daarbij waren onze overleggen ook nog eens gezellig. Dat alles lijkt misschien vanzelfsprekend, maar ik heb genoeg andere verhalen gehoord om te weten dat een goede begeleiding helemaal niet vanzelfsprekend is. Ik heb het erg goed getroffen.

Mijn ouders, Bert en Lottie van der Togt en Pien van der Togt bedank ik voor het feit dat ze mij indertijd hebben gestimuleerd om aio te worden, ook al dachten ze misschien dat het na een paar studies toch eigenlijk wel eens tijd werd voor een echte baan. Ik dank mijn familie en de familie Kloos voor de steun gedurende de afgelopen jaren en voor het feit dat ze nooit hebben gemopperd wanneer ik het weer eens liet afweten tijdens familiefestiviteiten. Van Eelkjen Kloos mocht ik bovendien een fragment van één van haar vele schilderijen gebruiken voor de voorkant van mijn boekje. Dank!

Aan Frank Bakker en Werner Brouwer heb ik indertijd bij BMG twee erg leuke kamergenoten gehad. Ik ben blij dat Frank en Bert Vrijhoef als ervaringsdeskundige vrienden bereid waren om tijdens de promotie bij mij te staan en mij bij te staan. Eindelijk kunnen Bert en ik onze afspraak nakomen om na het afronden van onze beider proefschriften in de Alpen te gaan fietsen. Bert heeft daarvoor lang op mij moeten wachten. In de Alpen zijn de rollen hopelijk omgedraaid.

Ik bedank ook mijn collega’s bij de LHV voor de steun. Een paar collega’s dank ik in het bijzonder. Karel Rosmalen heeft mij tijdens de laatste loodjes de ruimte gegeven om veel vrij te nemen. Christ Goossens kon vervolgens tijdens mijn vele afwezigheid voor mijn werk opdraaien, maar bleef altijd een gezellige en zorgzame kamergenoot. Met Annemarie Lamain heb ik gedurende het Switch-project veel samengewerkt. Het

afronden van mijn boekje viel deels samen met het maken van het Switch-boekje. Twee dingen tegelijk doen is niet mijn sterkste kant, maar kiezen ook niet. Ze bleef vrolijk en geduldig, ook wanneer ik niet meer wist hoe het verder moest. Marjolijn Rustemeijer bedank ik voor al die keren dat ze me heeft geholpen met het kopiëren en inbinden van tussentijdse versies.

Ten slotte bedank ik Ti. Haar bedank ik voor heel veel dingen. Zij heeft misschien nog wel het meest van iedereen uitgekeken naar de afronding van mijn proefschrift. Ik weet waarop ze zich in ieder geval enorm verheugt: een rituele verbranding van de stapels papier die veel te lang een beslag hebben gelegd op onze schaarse tijd en kastruimte. Verder ben ik erg blij dat ze mijn boekje heeft opgemaakt.

Mijn lieve Vita en Arthur bedanken voor hun belangrijke bijdragen aan mijn proefschrift is misschien wat overdreven. Toch noem ik ze hier. Het is vast leuk om één keer ’s avonds in bed uit mijn eigen boekje voor te lezen hoeveel ik van ze houd. Ik houd van jullie helemaal tot aan de maan – en terug!

CURRICULUM VITAE

Ad Vermaas (1965) attended secondary school in Zeist from 1977 to 1983. He then studied physiotherapy in Utrecht. After his graduation in 1988, he worked as a physiotherapist for nearly two years. From 1990 to 1994 he studied Health Policy and Management (BMG) at the Erasmus University Rotterdam. After graduation, he worked as a PhD student at the Department of Health Policy and Management (iBMG). Since 1999 he works as a policy assistant at the National Association for General Practitioners (LHV), where he contributed to the development and implementation of a new payment system for Dutch general practitioners.