Health status as a measure of outcome of disease and treatment

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Essink-Bot, Marie-Louise

Health status as a measure of outcome of disease and
treatment / Marie-Louise Essink-Bot. - Rotterdam:
Instituut Maatschappelijke Gezondheidszorg, Erasmus
Universiteit Rotterdam
Thesis Erasmus Universiteit Rotterdam. - With ref. - With
summary in Dutch.
ISBN 90-72245-64-4
Subject headings: health status ; measuring instruments /
quality of life ; measuring instruments.

Cover design and lay-out: I. Philips, Erasmus University
Printing: Drukkerij Humanitas, Rotterdam

Health status as a measure of outcome of disease and treatment

(Wat heet beter? Gezondheidstoestand als uitkomstmaat van ziekte en zorg)

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit
op gezag van de rector magnificus
Prof.Dr. P.W.C. Akkermans M.A.
en volgens besluit van het College voor promoties.

De openbare verdediging zal plaatsvinden op
woensdag 13 december 1995 om 11.45 uur

door

Marie-Louise Essink-Bot

geboren te Utrecht

Promotie commissie

Promotor:	Prof.Dr. P.J. van der Maas
Co-promotor:	Dr. G.J. Bonsel
Overige leden:	Prof.Dr. J. Passchier
	Dr. J. Kievit
	Prof.dr. F.H. Schröder

Dit proefschrift is mede tot stand gekomen dankzij de financiële steun van de Adviesgroep voor het Stimuleringsprogramma Gezondheidsonderzoek en van het instituut Maatschappelijke Gezondheidszorg, Erasmus Universiteit.

Contents

Descriptive health status measurement

Evaluative health status measurement

1

Introduction

1.1 Definition of the subject

The objective of medical care has always been twofold: to prolong life and to enhance its quality. Until recently, the emphasis was on prevention of premature death. An impressive increase in general life expectancy (partly as a result of technical progress in medicine) has caused a partial shift of emphasis, at least in the richer part of the world, towards the promotion of the quality of survival. This shift of purpose of medical care in the industrialized countries from curing acute, life-threatening disease to caring for patients with chronic diseases has created a different perspective on the balance of positive and negative effects of medical interventions. This can be illustrated with three examples.

Firstly, health status improvement, rather than prolonging life may be the primary aim of medical interventions nowadays, for example in the treatment of rheumatoid arthritis. The positive effects induced by relieving the pain should outweigh the possible negative side effects of the treatment (e.g., stomach complaints). If, in a second example, an intervention, for example treatment of cancer, is primarily intended to prolong life, a situation may occur in which the initial negative effects on health status effects must be weighted against the prolongation of survival. A third situation may be observed in emerging life-saving interventions, for example certain organ transplantations, the treatment of (otherwise lethal) congenital anomalies and of childhood cancer. The health status in the long term is gradually becoming the primary outcome criterion.

An outcome measure is needed to complement survival time in the evaluation of the impact of disease and of the effects of medical interventions. Quality of life measures are currently fulfilling this role. Assessment of the quality of life in this context is the subject of this thesis. It has been recognized that a person's quality of life is contingent upon various factors, such as a social network, the physical environment, financial situation, religion, as well as health. The first section of this Introduction has limited the notion of 'quality of life' in outcome assessment in the medical context to 'health-related quality of life', or 'health status', defined as quality of life relating to disease and/or treatment.

Health status measurement requires the application of principles from multiple disciplines, including the social sciences, medicine and economics. The input from the social sciences relates to methods of data-collection, instruments, and analytical techniques. Epidemiology has contributed to the field of health status measurement through the standard methodology for comparing medical interventions, i.e., the randomized controlled clinical trial. General principles of clinical epidemiology hold, irrespective of the outcome parameters being 'clinically' defined (e.g., tumour relapse) or in terms of 'patient outcome'. Because financial resources for providing health care have appeared to be limited, there was an obvious growing need for economic evaluation. '...No country in the world, even the richest, can afford to do all things that it is now possible to do to improve the health of their citizens. In that situation, it is no longer sufficient, in the competition for resources, solely to show that a particular intervention is beneficial, though that still is (or should be) a necessary condition for funding.'[1]

The perspective of this thesis, however, is medical, be it on an aggregate level. Information on patients' outcomes (including health status) can be useful to determine what is generally the best treatment for specific groups of patients. At an even higher level of aggregation, outcome data should be part of the information used in decision-making in the allocation of resources, for example at national level.

1.2 Objectives of the thesis

This thesis addresses a number of related topics in health status measurement in the evaluation of the effects of disease and of medical care. Its main objectives are:

1. To provide a general overview of the field of descriptive health status measurement.
2. To compare the contents and the relative performance of a number of currently available measures for descriptive health status measurement, to demonstrate applications of descriptive health status measurement and to discuss the importance of standardization of research methods in health status measurement.
3. To provide an overview of the current state of affairs in evaluative health status measurement, and to demonstrate empirical studies addressing the feasibility of collecting valuation data by self-assessment questionnaire.

1.3 Structure of the thesis

The thesis consists of four parts.

Part I (chapter 2) will provide the reader with a global overview of the field of health status measurement. This overview will include the relationship between conventional clinical parameters and health status measures, the distinction between descriptive and evaluative health status measurement, different perspectives on health status measurement and the consequences for research design, and finally arguments for standardization of research methodology in health status measurement and a description of the state of affairs with respect to such standardization in the Netherlands.

Part II (chapters 3 - 7) relates to descriptive health status measurement. Chapter 3 consists of an overview of the generic instruments that are currently available in the Netherlands, and of a comparison of these instruments based on users' opinions and the literature.

Chapters 4 and 5 are based on empirical studies on the relative performance profiles of a number of measures for generic health status assessment.

Chapter 4 contains a comparison of the Nottingham Health Profile and the Sickness Impact Profile when employed in a population of patients treated with renal dialysis.

In chapter 5, four measures (the MOS Short Form-36, the Nottingham Health Profile, the COOP/WONCA charts and the EuroQol instrument) are compared when employed in a study of migraine sufferers and a control group.

Chapters 6 and 7 present examples of applied health status measurement. Chapter 6 reports on a survey describing the health status of a representative sample of migraine sufferers when compared to that of a matched control group. Special attention is paid to the role of comorbidity. Chapter 7 presents the results of a longitudinal evaluation of the health status effects of liver transplantation.

Part III (chapters 8-10) relates to health state valuation research. Chapter 8 provides an introduction to the field of empirical valuation of health states. The chapter addresses issues such as the consequences of the current operationalization of the QALY-approach, in which health states are valued separately from their duration. Chapters 9 and 10 deal with empirical methodological research. Chapter 9 presents a pilot study investigating the feasibility of measuring valuations of health states among the general population in a postal survey. Chapter 10 attempts to analyse non-response and response behaviour in a survey aiming at the collection of valuations of health states in the general population.

Part IV (Chapter 11) presents a number of conclusions.

References

1 Williams A. *The importance of quality of life in policy decisions.* In: Walker SR, Rosser RM, eds. Quality of life assessment - key issues in the 1990s. Dordrecht: Kluwer Academic Publishers, 1993.

2

Overview of the field of health status measurement

Partly published as:
Essink-Bot ML, Bonsel GJ. Towards standardization of instruments for health status assessment (Naar standaardisatie van het instrumentarium voor het meten van de gezondheidstoestand) (in Dutch; English abstract). Huisarts en Wetenschap 1995;38(3):117-122. Reprinted with permission of the publisher.

2.1 Introduction

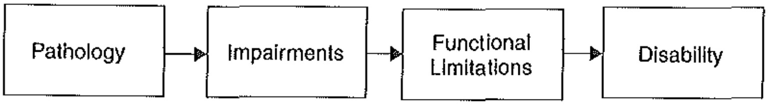
This chapter aims at providing a more detailed description of health status measurement. Firstly, the relationship between 'conventional' clinical parameters and health outcome measures will be clarified (2.2). In section 2.3 three main types of instruments for descriptive health status measurement will be introduced. The paramount distinction between description and valuation of health status is explained. Section 2.4 explains how these two concepts complement each other, and how their measurements are interrelated. Section 2.5 describes different perspectives on health status assessment and some of the consequences for research methodology. The position of 'time' in descriptive health status measurement is described briefly in section 2.6. The final section is dedicated to standardization of research instruments for health status assessment.

2.2 The relationship between conventional clinical parameters and health status measures

Nowadays the practice of health status assessment in medical evaluation research mainly relates to measuring the consequences of disease and/or treatment. The presence of disease may show itself at different levels, for example at the organ level, the level of the body system, and the level of a patient's functioning. The observation that medical diagnoses can be defined at different levels, e.g. at the level of the cell contents (e.g., a gene mutation), or at the level of disturbed functioning or well-being (e.g., fibromyalgia), may serve as an illustration. Distinguishing these levels of the consequences of disease and treatment implies that they have to be measured at these different levels. The level of interest determines which measure for health status is the most appropriate.

Conceptual schemes may provide an ordering of the different levels.[1,2] One of the available schemes, the 'Disablement Process', will be explained below. It is essentially similar to the International Classification of Impairments, Disabilities and Handicaps (ICIDH).[3]

FIGURE 2.2 The Disablement Process[1]



The first stage, 'Pathology' refers to functional or structural abnormalities at the level of the cell, for example, in biochemical processes or in genetic material. Detection often relies on evaluation of more manifest signs or symptoms. However, pathological abnormalities are increasingly detected directly as a consequence of technical diagnostic refinements. A specific class of 'pathology' has been emerging in recent years, i.e. genetically determined susceptibility to specific diseases.

'Impairments' are physiological and structural abnormalities at the level of organs or body systems (e.g., circulation). The consequences of atherosclerosis may become manifest through dysfunction of the arterial blood circulation of the brain, the heart or the legs. At the level of impairments, these consequences can be measured by clinical examination, laboratory tests, imaging techniques, in short, 'conventional' clinical measures.

'Functional limitations' are restrictions in performing 'generic' physical and mental actions (when compared with 'normal' performance by one's age group). Examples include walking, bending, seeing, hearing, speaking, thinking, laughing, crying etc. These are, as stated by Verbrugge et al, the basic interface between a person and the physical and social environment in which (s)he performs daily activities. Measures for functional limitations include performance tests and reports (by self or by a proxy) of performance of activities and the amount of difficulties experienced.

'Disability' represents difficulty performing activities in any domain of life due to a health problem. These domains cover for example work (paid and/or unpaid), household work, shopping, caring for children, hobbies, and travelling. Measuring disability requires respondents (or proxies) to answer questions on the difficulties experienced when performing specific activities. Alternatively, a subject's performance can be observed.

The relationship between the various stages of the Disablement Process is not straightforward. This was empirically supported by study results among stroke patients, which showed a pattern of decreasing correlations between 'stroke scales' (neurological impairments), the Barthel Index (functional limitations in the physical domain), the Rankin Scale (disability) and the Sickness Impact Profile (disability).[4] Pathology *may* lead to impairment (some abnormal cells turn out to be precursors of clinical cancer while others vanish). Impairment *may* lead to functional limitations (although high blood pressure does not necessarily do so). Functional limitations *may* lead to disabilities (the one patient returns to work after a myocardial infarction while another, with a

medically comparable infarction, is permanently disabled). Therefore, the relationship between the levels can more accurately be described in terms of probabilities.
Several types of factors, both intra individual and environmental, act as 'effect modifiers' along the way from pathology to disability. Examples of such factors are individual ability to adapt to illness (coping), environmental factors including housing situation and working situation, and factors intended to optimalize functioning, e.g., medical care, including rehabilitation.
The fact that there is no straightforward relationship between 'objective disease' (pathology, impairment) and the functioning of patients has consequences for the choice of health status measures. Clinical variables, including for example results of laboratory blood tests, imaging techniques and biopsies, and parameters focusing on a patient's functioning should be seen as complementary, each useful in their own context. 'Conventional' medical techniques can, for example, be used to determine the diagnosis of a disease, to decide what stage the disease is at, to support treatment decisions because they provide prognostic information, to monitor the course of the disease and to evaluate treatment effectiveness at a pathophysiological level.
Whether an intervention ultimately benefits a patient can be studied by evaluating patient's functioning in life, or 'patient outcome'. This applies when doctors treat individual patients as well as when medical interventions are evaluated at an aggregate level. Doctors should treat patients, not laboratory variables.

The concept 'patient outcome' is usually operationalized by two variables, i.e. survival time and health status (referring to the levels of 'functional limitations' and 'disability' in the Disablement Process). The concept and the measurement of survival are relatively straightforward. Health status, however, is a hypothetical concept without a direct empirical representation. The definition of domains is the first step to translation of such a concept into measurable terms. Comprehensive domains of health status current-ly include physical, psychological and social functioning. These domains should be subsequently operationalized to be measurable. It is agreed that the patient him/herself is generally the best source of information about his/her functioning in life. Data on a patient's functioning can be elicited by interview, self-assessment questionnaire or diary, or by observation. Due to the prevailing use of self-assessment questionnaires, the word 'measuring instrument' has become almost synonymous with 'questionnaire' in health status assessment. Although the following is restricted to self-assessment questionnaires for health status (with section 2.6 as an exception), this is not necessarily always the best method of data collection. Circumstances may occur in which self-assessment is not possible (for example, if the patient is a young child, or very ill, or illiterate).

2.3 Types of questionnaires for descriptive health status measurement

Three main types of questionnaires for health status assessment are distinguished: generic instruments, disease-specific instruments and domain-specific instruments.
Generic instruments for health status measurement are, by definition, comprehensive and non disease-specific. The items of such generic measures cover at least the physical, psychological and social domain in a non disease-specific way. The premise

underlying generic questionnaires for health status measurement is that different diseases have different consequences, but that these show themselves as different *patterns* of physical, psychological and social dysfunction. Generic questionnaires allow for comparison of health status data irrespective of diagnosis. The operationalization of the physical, psychological and social domains, as well as the choice of additional domains, may be different for different generic questionnaires. This may be explained by three reasons. Firstly, subtle differences in the concepts of health status that are used at the start of instrument development (e.g., more functionally oriented or more directed at perceptions of health status). Secondly, differences in the procedure of item selection, and thirdly different backgrounds of the researchers involved (psychology, sociology, economics, medical). Examples of generic instruments are the Sickness Impact Profile, the Nottingham Health Profile, the COOP/WONCA charts, the MOS-20, and the SF/RAND-36 (see Chapter 3).

Disease-specific instruments for health status measure the consequences of a *specific* disease or treatment. Treatment-specific measures for health status are often included in this group. Examples are cancer-specific instruments that contain detailed questions on the functional consequences of weight loss, hair loss, changes in body image etc; or instruments specifically designed for arthritis, with detailed questions on the consequences of joint pain, morning stiffness, etc. Many disease-specific instruments were developed in cancer research, but instruments have been developed for many other diseases: for example, arthritis, asthma, hypertension, and diabetes.[5,6,7,8,9,10,11,12,13,14]

Disease-specific measures usually show some overlap with generic measures. Some disease-specific measures may be viewed as narrowly focused generic instruments (e.g., measures designed for use with cardiovascular disease population or cancer populations).[15]

Disease-specific questionnaires often contain questions on symptoms and complaints. It is considered that questions regarding symptoms ['Were you short of breath?' (see note)] aim at measuring state of health at the level of 'impairment' (in this case the functioning of the respiratory system). Such formalized diagnostic questions, which aim at assessing whether disease or side-effects of treatment are present, and if so, how severe, should strictly not be seen as disease-specific measures *for health status* as operationalized in this thesis. If breathlessness is present, its impact on a patient's functioning still remains to be determined, see for example Wijkstra.[16] For example, a diagnostic questionnaire for arthritis could assess the degree of pain and morning stiffness, whereas an arthritis-specific questionnaire for health status assessment would ask to what extent the pain, the morning stiffness affect the patient's functioning. Complaints (e.g., 'Were you *bothered* by shortness of breath?'), as opposed to symptoms, contain an additional element of subjective experience.

Thirdly, *domain-specific instruments* measure the consequences of disease on a specific *domain* of health status, for example questionnaires on physical functioning, anxiety, depression, social relationships. Their use is not limited to a particular population. In this group of non-generic, non-disease specific instruments symptom-specific instruments can be seen as a separate group. They measure the effects of a specific symptom (e.g. pain, fatigue) on health status.

With respect to the context in which these three types of instruments may be most appropriately used, it may be said from a theoretical point of view that domain-specific health status measures (either physical, psychological or social functioning) are important but on their own rarely satisfactory. This is because a moderate degree of dysfunction in any of these domains usually has functional consequences for the others. If more than one domain may be expected to be affected, the choice should be between a disease-specific multi-domain questionnaire or a generic one. This issue will be discussed in more detail in section 2.5.

2.4 Description and valuation

Conventional descriptive health status scores take the form of a profile of scores across the different dimensions of the instrument. We have to go a step further if we wish to aggregate the consequences for health status and survival time into one outcome measure. Such a combined outcome measure is needed, for example, in cost-effectiveness analysis, assessment of the 'burden of disease' and in public health modelling.[17,18,19]

Suppose, for example, a measure has a physical dimension (A) with three levels (1=best, 2=intermediate, 3=worst) and a psychosocial dimension (B) with three analogous levels. Assume a patient's score profile is A_2B_2 while another patient's is A_1B_3. How can we judge whether the one patient is better off than the other? And if so, how much better?

To answer these questions we need to summarize the profile scores of health status. Such summary scores are currently obtained through a procedure in which health state descriptions are *valued*. A general overview of the field of evaluative health status measurement is presented in Chapter 8, but a short explanation of the current three-stage approach in empirical comprehensive outcome measurement is needed here as well.

In the first phase, actual patients' health states, as occurring in populations or resulting from an intervention, must be described in formal, functional terms. The EuroQol system for the description of health status (see Chapter 10) may be seen as an example. This system consists of 5 dimensions, each comprising 3 ordered categories, thus 3^5 (=243) health state descriptions are theoretically possible.

In the subsequent stage such health state descriptions are valued. The subjects who perform the valuation task are presented with a number of health state descriptions and are requested to rank these states according to the degree of (un)desirability and to indicate how good or how bad each of those states is for them.

In the third stage, the resulting value weights for health states are combined with survival data (e.g., into quality adjusted life years or QALYs). In public health research, population life years are often combined with quality data following the concept of 'Healthy Life Expectancy', in which a year lived in perfect health is valued '1' and a life-year with impairment (however defined) as '0'. By combining life-years with empirically collected values on health states the dichotomous value system of healthy life expectancy can be refined.[20,21,22]

Important choices in the empirical collection of valuation data to be addressed in this chapter have to do with the descriptive system for health status and the subjects who perform the valuation task. For other issues the reader is referred to Chapter 8.

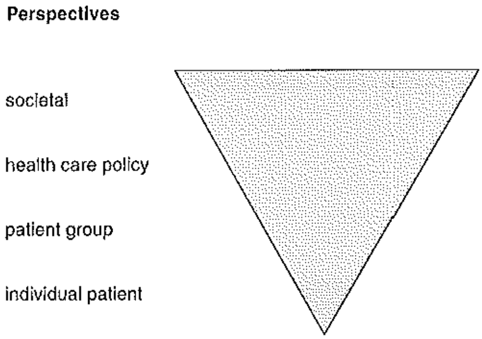
2.5 Different perspectives on health status assessment and their consequences

Health status is generally measured in order to support a decision-making process. Information on (expected) survival benefits may be important at each level of decision-making. However, this should be complemented by (different types of) health status information.
As recognized by Sutherland and Till, health status information and the levels of decision-making (perspectives) interact: the level determines which type of measuring instrument is the most appropriate.[23] In their paper, 3 levels of decision-making are distinguished, i.e. the micro (clinical) level, the meso (agency, institutional or regional) level and the meta (governmental) level. We will elaborate on the structure proposed by Sutherland and Till in the following by simplifying the distinction to essentially two levels of decision-making and relating the levels to methodological aspects of description and valuation of health status.

The *societal perspective* may be found at the one end of a hypothetical scale of decision levels (see Figure 2.5.1). Decisions in the public domain concern the distribution of funds over areas such as education, housing, public transport and health care. The societal perspective raises questions which relate for example to the relative benefits of investments in education and health care.
The *individual patient's perspective* may be found at the other end of the scale. Decisions to be taken from this perspective relate to the choice of the best treatment for the individual patient. Information on the health status of an individual patient is useful for the patient and the physician to guide such decisions.
Two perspectives are in-between. The first is the *patient group perspective*, which is essentially a higher level of the individual patient's perspective. Information on the relative benefits of two treatments for a circumscript group of patients may be used to guide decisions on the treatment to be preferred for that group of patients. In fact, it provides information on the health status effects to be expected for the individual patient. Research from the patient group perspective is referred to as medical evaluation research and the classical research design is the randomized controlled trial (RCT). Costs are not taken into account, at least not in analytical connection with effect measurement. This is because, as stated by Detsky, individual clinicians are appropriately concerned solely with the effectiveness of a specific intervention for their patients and are not concerned with the benefit derived from spending these resources on other patients.[24]

The second perspective which falls in-between the ends of the scale, the *health-care policy perspective* is essentially a narrowing of the societal perspective. Within the field of health care, decisions have to be taken regarding the distribution of funds, at a regional, local or institutional level. Research questions from this perspective relate for example to the relative effects of a screening programme for prostatic cancer and a programme to treat high blood cholesterol in middle-aged men with a drug. Evaluation research from the health care policy perspective is referred to as medical technology assessment (MTA). In MTA, costs and effects of different medical interventions are compared. In order to support decision-making in priority setting in health care, costs and effects of an intervention have to be compared with the costs and effects of other interventions for the same patient group, as well as with costs and effects of interventions for other patient groups. The methodology of economic evaluation, an important part of MTA, consists largely of the implementation of economic research methods alongside RCTs.[24,25,26,27]

FIGURE 2.5.1 Perspectives on health status assessment

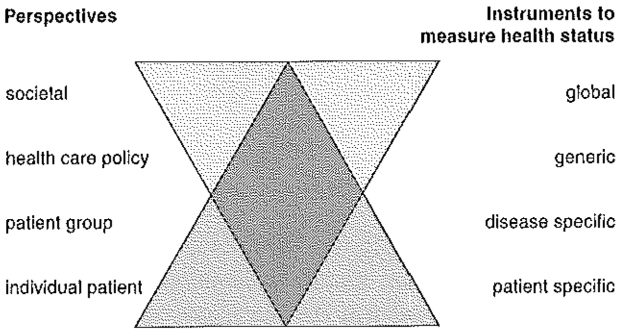


Nowadays medical evaluation research and MTA have become closely related. Due to the apparent scarcity of health care resources combined with an apparently infinite demand, it is generally considered that the 'no matter what it costs' assumption of RCTs[27] has a limited scope. The first question raised when evaluating new medical treatment is whether it is better overall than the conventional treatment. If this question is answered positively, i.e. a new treatment has a net benefit over the old one, this generally raises the question as to whether the difference is worth the difference in costs. The consequence is that MTAs and clinical effectiveness studies are often performed in close connection. This adds special requirements to the research design.

The level of decision-making (the perspective of a study) determines which type of measuring instrument is the most appropriate for description of health status (see Figure 2.5.2). For decisions at the *individual patient level*, the physician (and the patient) is interested in individual functioning and well-being. If a formal measuring instrument is used to assess health status at this level, it should be patient-specific. This ensures that the extent perceived by the patient of symptoms or side-effects interfering with his/her functioning is assessed. An overview of instruments for individual health status assessment is beyond the scope of this thesis.[28,29]

At a *patient group level* the aim is to gain insight into the functioning and the experiences of groups of patients with specific disease characteristics. Questionnaires must enable the description of the group under study and the comparison of results with those of a control group: a group of patients with comparable (disease and other) characteristics. It is especially important that the domains that are relevant to the functioning of patients with that disease are covered in detail. For example, in a study comparing breast conserving therapy and mastectomy in early breast cancer, the inclusion of body image questions is desirable as it has been found to be relevant for the functioning and well-being of these patients.[30] Disease- and treatment-specific instruments were designed for use in this context. However, by using a disease-specific instrument that covers only the domains that are expected to be affected by the disease and the treatment under study, changes in other domains may be overlooked. Therefore more general domains should be added to disease-specific ones in studies conducted from the patient group perspective.

FIGURE 2.5.2 Perspectives on health status assessment and their consequences for the choice of instruments



At the third level, i.e. the *health-care policy level*, the impact of different diseases and of different interventions on patient outcome should be compared. For example, comparing the effects of a screening programme for breast cancer to the effects of treating the clinically manifest disease, or comparing the effects of the screening programme to the effects of a liver transplantation programme.[31] The measures used to enable such comparison should be non disease-specific and comprehensive, in other words, generic.

As stated above, clinical trials (medical evaluation research) and MTAs are often conducted alongside each other.[24,26,27] The general recommendation is to employ a combination of a generic instrument complemented with one or more disease- and/or domain-specific instruments in medical evaluation research and in MTA. The generic results can be used to define the position of the patient group under study on the continuum ranging from the 'worst imaginable health state' to the 'best imaginable health state', and to relate the size of a treatment effect to this continuum. This is useful for comparisons with the effects of other interventions in similar or different patient groups. Results from the disease-, treatment- domain-, and symptom-specific instruments are useful to focus on characteristics of special interest, for example from a clinical point of view, in a particular patient group.

At the fourth level, the *societal perspective*, aspects other than health that contribute to overall quality of life are explicitly taken into account. The definition of 'quality of life' at this level has to be extended from the definition as used at the health care policy level and the patient group level, where 'non health-related' aspects of the quality of life are deliberately neglected. Respondents are asked about their health, education, housing, public transport, etc. in population based surveys. Obviously, health, as far as societal policy is concerned, is one issue among many others. With respect to health, a global score is generally asked for. Policy makers may use this information to obtain a domain specific and an overall view of the well-being or the level of (dys)-function of the population at large and of population subsets. Commonly used measures are those developed by Andrews & Withey and Campbell and colleagues to assess the quality of American life.[32,33]

Although the surveys are often named differently, most western countries have a nationwide survey to study the quality of life of their populations. Discussing health status assessment from the societal perspective is beyond the scope of this thesis.

The perspectives on health status assessment as distinguished above also have consequences for the valuation stage, for example for the choice of the subjects who perform the valuation task.

From the *individual patient perspective* the first question raised is, who should be involved in the decisions to be taken. Not every patient wants to take the lead or even to participate in the decision-making process.[34,35] If not, the treating physician may decide. The best decision is the one that most reflects the individual patient's values without being in conflict with values held by the physician. The patient's values do not necessarily have to be made explicit if (s)he is able to make his/her own decision. In cases where the patient wants to participate in the decision-making process but needs support, decision aids such as the board developed by Levine[36] may help the patient to become aware of his/her values with respect to the consequences of the different possible options and express them. The relevant information on the possible outcomes must come from data gathered in descriptive studies at a patient group level and from the clinician's clinical experience.

Data describing the outcomes of different interventions among similar groups of patients must first be available, preferably from randomized clinical trials, before deciding upon a policy regarding groups of patients with similar disease characteristics (the *patient group perspective*). Clinicians may choose the treatment policy generally to be preferred without explicitly weighting the different consequences of different lines of action. However, decisions can nowadays be supported by decision analytic approaches. For example, decision analysis was applied in the evaluation of various follow-up schedules for patients with colorectal cancer after intentionally curative surgery.[37] For decisions on treatment strategies within a disease category, values for the different possible outcomes by representatives of the patient group involved may be used.

In a decision-making process at a *health-care policy level*, the relative effects of different kinds of programmes (e.g., preventive versus curative) for different disease groups (e.g., cancer versus heart disease) within different age- and sex groups are compared in order to support the decision-making process in resource allocation. Implications for the valuation procedure are multiple. The values to be used should reflect the societal viewpoint, which is often operationalized by obtaining a representative sample from the general public, in their capacity as tax payers and as future patients, to perform the valuation task.[38]

Values of health states, combined with life-years form the basis of QALY calculations. The research field of evaluative health status measurement is complex. However, choosing to determine explicitly a set of relative values for different health outcomes and constructing a single outcome indicator provides opportunities to opening the 'black box' of the decision-maker's relative values for public scrutiny and influence.[22]

The experiences from the Oregon Medicaid Experiment show that the meaning of societal ratings on the desirability of health states ('values') needs careful explanation.[39] Oregon proposed a social experiment in which combinations of medical conditions and treatments were prioritized. The experiment intended to include empirical valuations of health states as a basic element. The US Department of Health and Human Services (DHSS) rejected Oregon's application to proceed with the experiment because the Oregon preference survey 'quantified stereotypic assumptions about persons with disability'. The DHHS assumed that health preference data were discriminatory because the health states of people without disabilities and of those with disability would not be rated as the same.[39] This statement shows clearly that the aim of the experiment was seriously misinterpreted and disclaims the notion that subjects who are at optimal health may need fewer health services than those who are less healthy.

2.6 The position of 'time'

'Time' occupies a special position in health status measurement. In the descriptive stage, time plays a role in the reference period of questionnaires (e.g., 'Think of the past day / week / month ...') and in the timing of assessments. In the valuation stage, time is incorporated for example in the duration of the state to be valued. Time occurs explicitly in the final stage of outcome evaluation when health status valuation data are combined with survival data. The issue of time in descriptive health status assessment will be addressed below, while 'time' in evaluative health status measurement is addressed in chapter 8.

The aim of descriptive health status measurement is to describe the course of health status over time. For each individual within a group of patients this may be operationalized by repetitive assessments that each refer to a defined period of time during which health status is assumed to be stable. This implies that the results of each assessment should be globally representative of the distribution of health status during a given period. We are generally not interested in a unique point estimator. Therefore, many questionnaires use a reference period; for example, the COOP/WONCA charts refer to 'the past two weeks', the EORTC QLQ-C30 to 'the past week', the SF-36 to 'the past four weeks', the SIP to 'today'. The time-frame of a questionnaire should theoretically be in accordance with the length of the period for which the assessment is assumed to be representative. However, from a psychometric point of view, short time frames are preferable if health status ('state') is to be differentiated from personality traits including complaint behaviour. This was illustrated by the results of a study that showed that responses of subjects about their complaints with 'yesterday' as a time frame could be clearly differentiated from their answers to questions relating to neuroticism (a personality trait). When this time span was enlarged from 1 to 3 days, responses to the complaint items could not be clearly differentiated from the measure of neuroticism.[40]

Establishing that assessments should be representative of periods of stable health status also has consequences for the number and timing of assessments. The commonly applied model of one assessment before and one assessment after an intervention is only justified if the individual shows a stable health status before the intervention, a reaction on treatment that is restricted to a well-defined period and a stable health status afterwards. The reaction on treatment must be homogeneous in direction and in magnitude for all domains of health status. Moreover, the group of patients must be homogenous with respect to pre-intervention level of health status, reaction on treatment and post-intervention level (see Figure 2.6.1).

In practice cases are commonly characterized by:

- different levels of health status among individuals before the intervention (see Figure 2.6.2);
- heterogenous reactions on treatment [referring to either inter-individual differences in the speed of the reaction (see Figure 2.6.3), or intra-individual differences in the reaction on treatment for different domains of health status (see Figure 2.6.4)];
- different courses among individuals after the intervention [including for example, uncomplicated recovery, complicated recovery, and a declining course resulting in death (see Figure 2.6.5)].

General guidelines for the timing of assessments are difficult to give, but the commonly observed once before - once after model for data analysis is seldom valid.[41]

FIGURE 2.6.1 Course of health status over time (i): stable health status before treatment; homogenous reaction on treatment

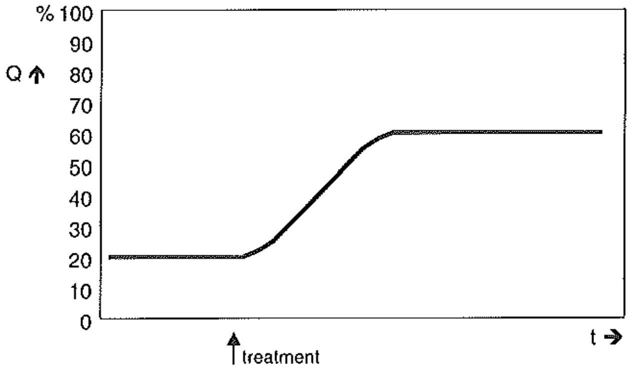


FIGURE 2.6.2 Course of health status over time (ii): inter-individual differences in level of health status; homogenous reactions on treatment

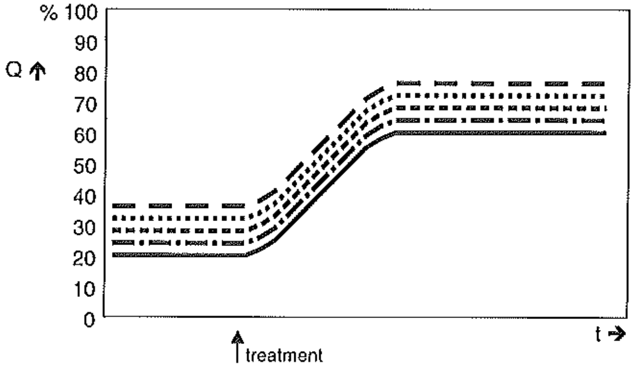


FIGURE 2.6.3 Course of health status over time (iii): inter-individual differences in the speed of the reaction on treatment; homogeneity in magnitude and direction of the reaction

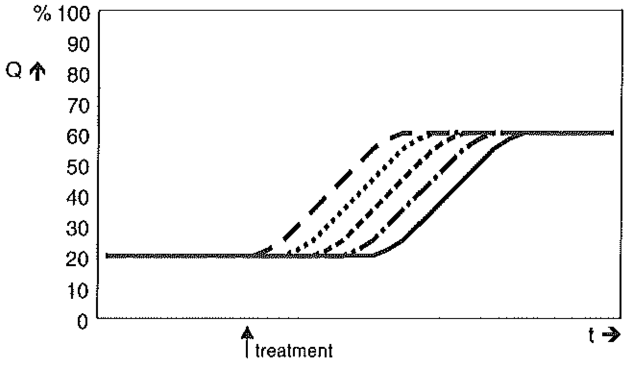


FIGURE 2.6.4 Course of health status over time (iv): intra-individual differences in the reaction on treatment for different domains (Q) of health status

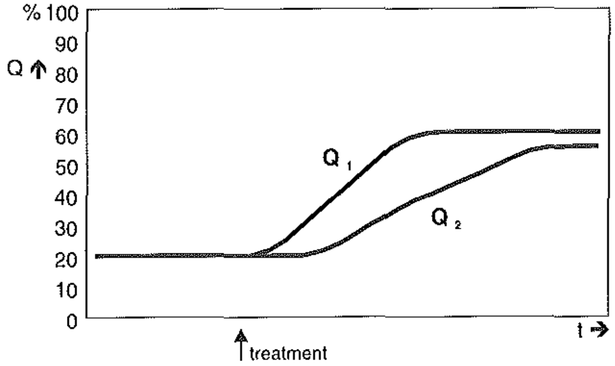
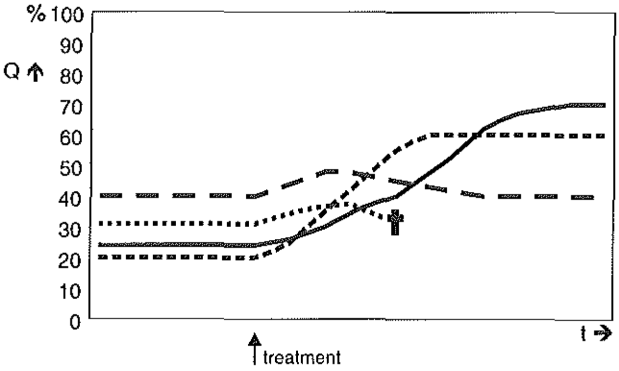


FIGURE 2.6.5 Course of health status over time (v): inter-individual differences in health status before treatment and in direction, speed and magnitude of the reactions on treatment



2.7 Standardization of research instruments for health status assessment

Generally, methodological choices in designing a research project depend on the research question to be addressed. However, as explained below, there are arguments to strive for some level of standardization of research methodology.

As stated in a section 2.5, the perspective (i.e., the level of decision-making at which the information is to be used) determines the research question in health status assessment, and consequently partly the research methods to be employed. If health status is measured from the health care policy perspective, which is a societal perspective, comparability is an essential part of the research question. The meaning of 'comparability' in this context is twofold. Firstly, it refers to comparability across diagnoses and disease stages, requiring the use of generic health status measures. Secondly, comparability means a comparison with other studies that include health outcome measures. The latter requires much further standardization of the complete design of evaluation studies, including the choice of generic questionnaires, the timing of assessments, the choice of valuation methods, the methods of data analysis, presentation of the results, etc. In RCTs conducted from the patient group perspective, comparability of results with other studies may be not strictly essential. However, it is considered efficient to additionally strive for some comparability in research from the patient group perspective.

The way to achieve a higher degree of comparability of study results is to define and implement a certain level of standardization. This implies some restriction of the individual freedom of researchers and research groups.

The remainder of this chapter relates to standardization of generic instruments for health status assessment. Other aspects of research design that might be related to standardization, although equally important with respect to comparability of study results, are not discussed.

2.7.1 *Levels of standardization*

The prescription of one standard measuring instrument is the most rigid level of standardization that could be imagined. The present situation may be seen as the other, liberal, end of a hypothetic continuum. The choice of measuring instruments by the individual researcher or research group is presently based on a multitude of heterogenous considerations, including the research question, personal taste, tradition, fashion, and investments made by using an instrument in previous reseach.

Assuming a foreign origin for most instruments for health status assessment, levels of standardization can be ranked as below.

1. Standardized rules for the adaptation of foreign instruments; free choice from the available instruments.
2. One standardized version for each instrument; free choice.
3. One standardized version for each instrument; restriction of freedom of choice by recommending to choose at least one from a set of questionnaires.
4. One standardized version for each instrument; recommendation to include a standard 'common core' instrument.
5. One standard instrument; no other options.

Reaching levels 1 or 2 would imply an improvement compared to the present situation, although this would not contribute very much to comparability of the results of different studies employing different health status measures. It is shown in chapter 3 that the generic instruments currently available in the Netherlands are rather different, precluding simple 'translation' of results from one instrument to another. Level 5 is of course undesirable as it would destroy all creativity, consequently hampering scientific progress.

Level 4, representing the 'common core' option, apart from standardization of *versions* of questionnaires, is desirable and might seem to be feasible as well. Such a common core can be seen as a minimal level of standardization, providing at least the possibility to compare between patient groups and to locate these groups on the hypothetical continuum between 'the best imaginable health state' and 'the worst imaginable health state'. The development of the EuroQol instrument emanated from the 'common core' principle.[42]

2.7.2 Ways towards standardization

Several ways towards standardization, each differing in the amount of active effort required, can be discerned. The first is 'wait and see', as in the long run, the 'best' instrument will gain general acceptation and become the standard. The disadvantages of this passive strategy include the long time the waiting may take and the fact that factors other than scientific quality (for example, effective marketing) may determine the dissemination of an instrument.

A second way is also passive. 'Creeping standardization' describes the process of copying of methodology without further reflection or testing of instruments. An instrument may be chosen automatically for use in a research project because everybody has used it, despite the lack of empirical research undertaken on the relative qualities of the instrument.

A third strategy includes an active marketing approach. Given sufficient financial resources are available, an instrument may be developed, tested and become very popular in a short time, as has been demonstrated by the introduction of the SF-36.

A fourth strategy provides researchers with guidelines as to the choice of instruments, which are based on scientific information and consensus reached among experts in the field. This strategy, which seems the most desirable, is being followed by the Dutch Working Group on Health Status Assessment, see the next section.

2.7.3 Implementation of standards: Dutch Working Group on Health Status Assessment (Werkgroep Onderzoek Gezondheidstoestandmeting)

In the context of the research programme 'Standardization in Medical Technology Assessment', funded by the Health Research Promotion Programme (Stimuleringsprogramma Gezondheidsonderzoek), a group of experts representing the field of health status assessment in the Netherlands have met at regular intervals in the Dutch Working Group on Health Status Assessment (Werkgroep Onderzoek Gezondheidstoestandmeting) since 1992. The aims of the Working Group are, firstly, to promote the use of health status measures in clinical research; secondly, to bring about a certain level of standardization in order to increase the comparability of study results; thirdly, to establish a nationwide network for researchers engaged in health status measurement. As consensus appeared to be feasible on a number of topics considered essential for comparability, the Working Group intends to provide researchers with guidelines based on that consensus and the scientific 'state of the art'.

Important guidelines reached by the middle of 1994 were:

1. *When to include health status assessment in clinical research.* The Working Group, following a recommendation of the National Cancer Institute of Canada[43], recommends to include a paragraph on health status assessment in all proposals for

medical evaluation research. This paragraph should explain why health status measurement is or is not included in the study.

2. *Generic measuring instruments.* The Working Group recommends data collection for reference purposes with a 'common core' instrument in any medical evaluation project. The COOP/WONCA charts are recommended as the 'common core' instrument for 1994-1996. During this period the usefulness of the 'common core' and its contribution to the comparability of study results will be evaluated. Additionally, reliability and validity studies of the COOP/WONCA charts will be continued in a wide range of patient populations.

Additionally, the Working Group intends to promote methodological research in the field of health status assessment and to collect and disseminate structured information on health status assessment. Information and guidelines will be disseminated by seminars, publications in the scientific press[44] and an instruction booklet for clinical researchers[45].

Note

1. This is an item from the EORTC QLQ-C30.

References

1 Verbrugge LM, Jette AM. *The disablement process.* Soc Sci Med 1994;38:1-14.

2 Wilson IB, Cleary PD. *Linking clinical variables with health related quality of life.* JAMA 1995;273:59-65.

3 World Health Organization. *International classification of Impairments, Disabilities, and Handicaps.* Geneva, 1980.

4 Haan R de, Horn J, Limburg M, Van der Meulen J, Bossuyt P. *A comparison of five stroke scales with measures of Disability, Handicap and Quality of Life.* Stroke 1993;24:1178-1181.

5 Haes JCJM de, VanKnippenberg JCE, Neijt JP. *Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist.* Br J Cancer 1990;62:1034-1038.

6 Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. *The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.* J Nat Cancer Inst 1993;85:365-376.

7 Meenan RF, Gertman PM, Mason JH. *Measuring Health status in arthritis: the Arthritis Impact Measurement Scales.* Arthritis Rheum 1980;23:146-152.

8 Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ. *Evaluation of the Dutch Arthritis Impact Measurement Scales in patients with rheumatoid arthritis.* Brit J Rheumatol

1989;28:487-491.

9 Hyland ME, Finnis S, Irvine SH. *A scale for assessing quality of life in adult asthma sufferers.* J Psychosomatic Research 1991;35:99-110.

10 Juniper EF, Guyatt GH, Epstein RS et al. *Evaluation of impairment of health-related quality of life in asthma: develpment of a questionnaire for use in clinical trials.* Thorax 1992;47:76-83.

11 Rutten-van Mölken MPMH, Custers F, Doorslaer EKA van, Jansen MCM, Heurman L, Maesen FPV et al. *Comparing the performance of four different instruments in evaluating the effects of salmeterol on asthma quality of life.* Eur Resp J 1995; forthcoming.

12 Wijkstra PJ, Vergert EM Ten, Van Altena R, Otten V, Postma DS, Kraan J, Koëter GH. *Reliability and validity of the chronic respiratory questionnaire (CRQ).* Thorax 1994;49:465-467.

13 Fletcher AE, Chester PC, Hawkins CMA, Latham AN, Pike LA, Bulpitt CJ. *The effects of verapamil and propranolol on quality of life in hypertension.* J Human Hypertens 1989;3:125-130.

14 Grootenhuis PA, Snoek FJ, Heine RJ, Bouter LM. *Development of a Type 2 Diabetes Symptom Checklist: a measure of symptom severity.* Diabetic Medicine 1994;11:253-261.

15 Aaronson NK. *Quantitative issues in health-related quality of life assessment.* Health Policy 1988;10:217-230.

16 Wijkstra PJ, Vergert EM Ten, Van Altena R, Otten V, Postma DS, Kraan J, Koëter GH. *Reliability and validity of the chronic respiratory questionnaire (CRQ).* Thorax 1994;49:465-467.

17 Drummond MF, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes.* Oxford: Oxford University Press, 1987.

18 World Bank. *World Development Report 1993: investing in health - world development indicators.* Oxford: Oxford University Press, 1993.

19 Ruwaard D, Kramers PGN, eds. *Volksgezondheid Toekomst Verkenning.* Den Haag: SDU, 1993.

20 Mackenbach JP. *The future of public health in general, and 'healthy life expectancy' in particular.* (De toekomst van de volksgezondheid in het algemeen, en de 'gezonde levensverwachting' in het bijzonder; in Dutch). Ned Tijdschr Geneeskd 1994;138(21):1053-1055.

21 Barendregt JJ, Bonneux L. *Changes in incidence and survival of cardiovascular disease and their impact on disease prevalence and health expectancy.* (De invloed van veranderingen in incidentie en overleving van cardiovasculaire ziekten op ziekteprevalentie en gezonde levensverwachting; in Dutch, English abstract). Tijdschr Soc Gezondheidszorg 1994;72(8):429-433.

22 Murray CJL. *Quantifying the burden of disease: the technical basis for disability-adjusted life years.* Bulletin of the WHO, 1994;72(3):429-445.

23 Sutherland HJ, Till JE. *Quality of life assessments and levels of decion making: differentiating objectives.* Quality of Life Research 1993;2:297-303.

24 Detsky AS, Naglie IG. *A clinician's guide to cost-effectiveness analysis.* Ann Int Med 1990;113:147-154.

25 Bonsel GJ. *Medical Technology Assessment crossing your path.* (Als Medical Technology Assessment uw pad kruist, in Dutch).Tijdsch Soc Gezondheidszorg 1992;70:487-489.

26 Drummond MF, Davies L. *Economic analysis alongside clinical trials.* Int J TA in Health Care 1991;7:561-573.

27 Bonsel GJ, Rutten FFH, Uyl-de Groot CA. *Measurement and valuation of quality of life in economic appraisal of cancer treatment.* Eur J Cancer 1994;30A(1):111-117.

28 O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CRB. *Individual Quality of life in patients undergoing hip replacement.* Lancet 1992;339:1088-1091.

29 McGee HM, O'Boyle CA, Hickey A, O'Malley K, Joyce CRB. *Assessing the quality of life of the individual: the SEIQOL with a healthy and a gastroenterology unit population.* Psychol Med 1991;21:749-759.

30 Kiebert GM, Haes JC de, Velde CJ van de. *The impact of breast-conserving treatment and mastectomy on the quality of life of early-stage breast cancer patients: a review.* J Clin Oncol 1991;9(6):1059-1070.

31 Haes JCJM de, Koning HJ de, Oortmarssen GJ van, Agt HME van, Bruyn AE de, Maas PJ van der. *The impact of a breast cancer screening programme on QALYs.* Int J Cancer 1991;49:538-544.

32 Andrews FM, Withey SB. *Social indicators of well-being.* New YorK: Plenum Press, 1976.

33 Campbell A, Converse PE. *The Quality of American life.* New York: Russell Sage Foundation, 1976.

34 Sutherland HJ, Llewellyn-Thomas HA, Lockwood GA, Tritchler DL, Till JE. *Cancer patients: their desire for information and participation in treatment decisions.* J Royal Soc Medicine 1989;82:260-263.

35 Blanchard CG, Labrecque MS, Ruckdeschel JC, Blanchard EB. *Information and decision-making preferences of hospitalized adult cancer patients.* Soc Sci Med 1988;27(11):1139-1145.

36 Levine MN, Gafni A, Markham B, McFarlane D. *A bedside decision instrument to elicit a patient's preference concerning adiuvant chemotherapy for breast cancer.* Ann Int Med 1992;117(1):53-58.

37 Bruinvels DJ. *Follow-up of patients with colorectal cancer.* Dissertation, Leiden, The Netherlands, 1995.

38 Hadorn DC. *The role of public values in setting health care priorities.* Soc Sci Med 1991;32:773-781.

39 Kaplan RM. *Value judgment in the Oregon Medicaid Experiment.* Med Care 1994;32:975-988.

40 Huisman SJ, Van Dam FSAM, Aaronson NK, Hanewald GJFP. *On measuring complaints of cancer patients: some remarks on the time span of the question.* In: The quality of life of cancer patients. NK Aaronson, J Beckman, eds. New York: Raven Press, 1987.

41 Bonsel GJ. *Methods of medical technology assessment with an application to liver transplantation.* Thesis. Rotterdam, 1991.

42 EuroQol Group. *EuroQol - a new facility for the measurement of health-related quality of life.* Health Policy 1990;16:199-208.

43 Osoba D. *The quality of life committee of the clinical trials group of the National Cancer Institute of Canada: organization and functions.* Quality of Life Research 1992;1:211-218.

44 Essink-Bot ML. *The Dutch Working Group on Health Status Assessment; standardisation of health-related quality of life research.* (De Werkgroep Onderzoek Gezondheidstoestand-meting; standaardisatie van onderzoek naar met gezondheid samenhangende kwaliteit van leven; in Dutch). Ned Tijdschr Geneeskd 1994;138:1484-1486.

45 Essink-Bot ML, Haes JCJM de. *A short guide to quality of life in medical research.* (Kwaliteit van leven in medisch onderzoek - een korte inleiding; in Dutch). In preparation.

3

An overview of six generic instruments for health status assessment

Partly published as:
Essink-Bot ML, Bonsel GJ. Towards standardization of instruments for health status assessment (Naar standaardisatie van het instrumentarium voor het meten van de gezondheidstoestand) (in Dutch; English abstract). Huisarts en Wetenschap, 1995; 38(3):117-122. Reprinted with permission of the publisher.

3.1 Introduction

Generic instruments for health status are, by definition, comprehensive and non-disease specific. An instrument is defined as 'comprehensive' if at least the physical, the psychological and the social domains are covered. The items of a generic health status instrument cover these domains in a non disease-specific way. This combination of characteristics makes generic instruments especially suitable for comparison of study results between different disease stages and diagnostic groups.
The present chapter consists of a comparison of the contents and testing properties of six generic questionnaires that are currently available in the Netherlands. The comparison, presented below, was based on the available literature, own research and on expert judgements of the users of the questionnaires.

3.2 Six generic instruments for health status assessment

Six generic questionnaires for health status assessment available and commonly used in the Netherlands will be introduced and compared below. The selection includes the Sickness Impact Profile (SIP), the Nottingham Health Profile (NHP), the COOP/WONCA-charts, the 20/24-item instrument from the Medical Outcomes Study (MOS-20/24), the MOS 36-item Short Form Health Survey (SF-36) / RAND 36-item Health survey 1.0 (RAND-36) and the EuroQol instrument. References refer to Dutch adaptations of the questionnaires.
The *Sickness Impact Profile* was developed in the US between 1972 and 1981 in order to assess the consequences of disease and treatment in functional terms. Items were selected from a pool of statements describing sickness-related changes in behaviour. These statements were obtained from patients, health-care professionals, individuals caring for patients and apparently healthy subjects.[1] The 136 items belong to 12 scales (see Table 3.2). Apart from a 12-dimensional profile score, the SIP is capable of generating a 'physical' subscore (3 scales), a 'psychosocial' subscore (4 scales) and a total score. The time-frame of the SIP-items is 'today'. Examples of SIP items are shown in Chapter 4. In the self-assessment version of

the SIP, the respondent is requested to tick the statements that apply to him/her in relation to his/her health. The SIP was adapted for use in the Netherlands by researchers of the Utrecht Institite for General Practice. In the Dutch situation the SIP has been extensively used in rehabilitation studies, neurology, and in general practice. A shortened SIP (68 items) with comparable qualities has recently become avalaible.[2,3,4,5]

The *Nottingham Health Profile* was developed during 1977-1981 in Great Britain as a measure for perceived health. Items were selected from an item-pool that was created by interviewing patients with a variety of chronic ailments.[6] The NHP was intended for use in population surveys. It consists of two parts. Part 1 consists of 38 dichotomous items, covering the scales as listed in Table 3.2. Examples of NHP-1 items are shown in Chapter 4. Part 2 consists of seven items on problems because of health in specified areas of life (for example, paid employment, household work, social life). The time reference period in the NHP is 'at the moment'. Several Dutch NHP versions are available. The NHP-Dutch Adaptation by Bonsel et al. has been tested in several patient populations. An agreement was reached with the authors of the other Dutch version of the NHP to strive for one documented Dutch NHP.[7,8]

De *COOP/WONCA-charts* were developed to assess health status of patients in primary care. Several well-established instruments were used as a source for the contents of the charts. There are 6 charts covering the domains mentioned in Table 3.2. A Pain chart is optionally available. Each item has five function levels which are illustrated with pictogrammes. These pictogrammes may be useful aids for groups with low reading ability and/or low mastery of the Dutch language. The time-frame refers to the 'past two weeks'. There is one Dutch standard version.[9,10]

The *20/24-item instrument from the Medical Outcomes Study* is a summary version of the RAND Health Insurance Study Questionnaire which was used in the US in the RAND Health Insurance Experiment to study the health status effects of types of health insurance. In the Medical Outcomes Study framework of health indicators the two theoretical dimensions of health, i.e., physical and mental, are defined in terms of a variety of indicators. After a content area was specified, items were written to operationalize each concept.[11] The content of the MOS-20 was chosen to represent only the most important health concepts. The items of a dimension were selected to meet minimum standards of precision for purposes of group comparisons.[11] The time reference periods in the items of the MOS-20 are one month or three months. An 'acute version' with shorter reference periods is available. The 20 items cover 6 scales (see Table 3.2).[12] The most important difference between the two existing Dutch versions [by Kempen (NCG, Groningen) and De Haes (AMC, Amsterdam), respectively] is the addition of 4 items on Vitality in de latter version, making it a MOS-24.

The *MOS 36-item Short Form Health Survey / RAND 36-item Health survey 1.0* is a 36-item version of the same instrument. The longer 36-item version was developed to improve some shortcomings that were observed when employing the MOS-20.[13] For example, a floor phenomenon was found, i.e., a lower ability of the MOS-20 to discriminate among health states of seriously ill patients.[14] Two US originals exist, which are so similar that we regard them as identical. The same applies to the two co-existing Dutch versions, which are referred to as the SF-36 and the RAND-36 respectively. Any further reference to the SF-36 in the following will also apply to the RAND-36. The 36 items cover 8 scales (see Table 3.2). The item of Reported Health Transition is scored separately.[15,16] An example of a SF-36 item is: 'During the past 4 weeks, how much of the time has your physical health or emotional problems uinterfered with your social activities (like visiting with friends, relatives, etc.)?' An 'acute version' with a reference period of 1 week is available.

The *EuroQol classification* consists of five items (see Table 3.2). The choice of the dimensions was guided by a careful review of existing descriptive health status measures. Each item comprises the following levels: no problems - some problems - extreme problems. Additionally, evaluation of own health is assessed with a visual analogue scale ('thermometer') ranging from 0 (worst imaginable health state) to 100 (best imaginable health state). The time-frame of the EuroQol instrument is 'today'.

EuroQol health state descriptions can be linked directly to empirical valuations of health states by the general public, a feature which makes it especially interesting for use in economic evaluations of health care interventions.[17] The EuroQol instrument was developed by the international EuroQol Group and is intended 'to complement other quality-of-life measures and to facilitate the collection of a common dataset for reference purposes'.[18] There is one Dutch standard original available.

3.3 Qualitative comparison of the contents

A qualitative comparison of the item-content was carried out on the multi-item scales of the SIP, NHP, MOS-20/24, SF-36, and on the items of the COOP/WONCA charts and the EuroQol. Scales/items were considered comparable if their content was judged to refer to the same general health domain; see Table 3.2.

The physical domain is operationalized with an emphasis on walking (SIP-Ambulation, NHP, EuroQol) or as overall physical functioning (COOP/WONCA, MOS-20/24, SF-36). The SIP adds a dimension labeled 'Mobility' which relates to 'range of action', an issue that goes uncovered in the other five questionnaires. The psychological domain is similarly present in the six instruments. The same holds for role functioning, which is however somehow underrepresented in the NHP. Social Isolation$_{\text{(NHP)}}$ relates to the ability to make contact with other people and was considered to belong to the psychological rather than the role domain.

A pain dimension is absent in the SIP and in the COOP/WONCA charts (though optionally available in the latter). Pain$_{\text{(MOS-20)}}$, Bodily pain$_{\text{(SF-36)}}$ and Pain$_{\text{(NHP)}}$ do not refer to somatic sensations other than pain. Dimensions relating to overall health are available in 4 out of 6 questionnaires. Some instruments contain unique dimensions (Sleep$_{\text{(NHP)}}$, Alertness$_{\text{(SIP)}}$, Communication$_{\text{(SIP)}}$). The SF-36 and MOS-24 are the only instruments to address the concept of positive health (items worded as 'full of pep' in the Vitality-scale). NHP$_{\text{(Energy)}}$ is not an indicator of positive health as all items are phrased negatively ('tired all the time', 'everything is an effort', 'soon running out of energy').

This leads us to the following conclusions. Broadly speaking, all six instruments address two basic health domains, i.e., physical and psychosocial functioning. However, each approaches these domains areas from a somewhat different perspective. Similarity in scale labels sometimes hides dissimilarities in content. The reverse (similar content, dissimilar labels) also occurs.

TABLE 3.2 Qualitative comparison of questionnaire content of SIP, NHP, COOP/WONCA charts, MOS-20/24, SF-36, EuroQol

SIP	NHP	COOP/WONCA	MOS-20/24	SF-36	EUROQOL
--	--	Physical fitness	Physical functioning	Physical functioning	--
Ambulation	Physical Mobility	--	--	--	Mobility
Mobility	--	--	--	--	--
Emotion	Emotional reactions + Social Isolation	Feelings	Mental health	Mental health	Anxiety/ depression
Household management + Social interaction + Work + Recreation and pastimes	--	Daily Activities + Social Activities	Role functioning + Social Functioning	Role-physical + Role-emotional + Social functioning	Daily activities
--	Pain	--	Pain	Bodily pain	Pain/ discomfort
--	--	Overall health	Current health perceptions	General health perceptions	Valuation own health
Bodycare and movement + Eating	--	--	*	*	Self care
--	Energy	--	Vitality (24)	Vitality	--
Sleep and rest	Sleep	--	--	--	--
--	--	Change in health (2 weeks)	--	Reported health transition (1 year)	--
Alertness	--	--	--	--	--
Communication	--	--	--	--	--

* Item in the Physical Functioning scale

3.4 General criteria for the quality of health status measures

We discern three types of criteria for quality of health status measurement instruments, i.e., practical, technical and conceptual criteria. Each will be addressed below.

Practical criteria that determine the feasibility and consequentially the applicability of questionnaires, can be summarized as 'respondent burden'. 'Length', or the number of items of a questionnaire, represents only one aspect of respondent burden. Other aspects, including the degree of complexity and required reading ability should be taken into account as well. Moreover, the feasibility of a questionnaire is population specific. Empirical health status measurement is generally feasible even for seriously ill patients if good explanations about the aims and the necessity of the research are provided to patients, physicians and nursing staff, and provided that the whole procedure is extremely user friendly. Computer-assisted interviewing may offer special advantages, for example by providing invisible routing procedures.

Technical criteria relate to reliability, both in terms of internal consistency and of test-retest-, inter-observer-, and intra-observer-reliability (the latter two not applicable for self-assessment instruments). These are considered technical criteria because there is a general consensus about appropriate testing procedures and statistics (although the use of the intraclass correlation coefficient might be further encouraged).[19] An issue deserving special attention with respect to instruments for health status measurement is that reliability estimates obtained in one population may not be generalized to other populations with different characteristics regarding, for example, age, sex and disease-severity distributions.

Conceptual criteria relate to validity, or the extent to which the instrument measures the characteristic as intended. Three types of validity are distinguished in classical test theory: criterion validity, content validity, construct validity. Determination of criterion validity requires a measurable superior reference criterion, which is generally not available for health status measurement. Content validity refers to theoretical testing of the contents of an instrument (i.e., representative coverage of all relevant domains). Construct validity requires empirical testing of a priori hypotheses about the relationship of the instrument under study with instruments of proven validity. For health status instruments, this generally takes the form of testing the instrument's ability to discriminate among 'known groups' ('clinical validity') and of comparing data from the new instrument with simultaneously obtained data from other health status measures.

A special feature of validity testing of health status instruments is, again, its population specificity. An instrument which measures health status adequately in a relatively healthy population still remains to be validated in seriously ill populations. An aspect of validity deserving further research is responsiveness to (clinical) change over time, or 'longitudinal validity'.[20] This issue should be preceded by determination of test-retest reliability.[21] If a measure is known to be stable over time when health status did *not* change, it is useful to investigate if the instrument is able to reflect actual changes over time. The ability to discriminate between groups at one moment in time does not garantuee a good responsiveness over time.

Generally, validity testing should be seen as a continuous proces, that yields indications about the degree of confidence we can place on inferences that are made about people based on their scores from an instrument.[21]

3.5 Qualitative comparison of questionnaire properties

Some properties of the questionnaires are listed in Table 3.5. The data in the table represent combined literature data on the Dutch versions of the questionnaires and expert opinions from members of the Dutch Working Group on Health Status Assessment (see Chapter 2, section 2.7), all of them engaged in developmental work on Dutch adaptations of foreign questionnaires.

All 6 instruments are suitable for self-assessment. The number of items differs greatly. SIP, NHP (Part 1), MOS-20/24 and SF-36 are classical multi-item scales (although some scales contain only 2 or 3 items), while COOP/WONCA and EuroQol are classification instruments. The reported completion times probably relate to relatively well patients. Questionnaires appear to be tested in patient groups that can be approached relatively easy (e.g., not too seriously ill, not too old, no vision problems, suffering from chronic diseases with a fairly predictable course). Data on questionnaire behaviour in 'difficult' populations are scarce.

Different types of response choices (i.e., dichotomous, Likert) are applied. A consequent relatively easy response mode enhances the 'feasibility' of an instrument. The risk of acquiescence bias is the other side of this picture.[21] The occurrence of different response choices in one questionnaire probably adds to the degree of difficulty experienced by the respondents.

All 6 questionnaires meet the criteria (see section 3.1) for generic instruments. However, some of them were originally designed for groups of patients with rather specific diagnoses. The fact that the early research which eventually led into the development and testing of the present NHP was conducted in the context of hip replacement operations may explain the relative emphasis on walking.[6,22]

Applicability of questionnaires in different age groups implies fulfilling two distinct criteria. The first relates to the cognitive ability necessary for filling in a questionnaire. The second relates to the validity (including 'face validity', i.e., the appropriateness of the items) in different age groups.

We originally intended to include a comparison of psychometric testing properties of the six instruments (reliability, validity). The attempt to combine literature and expert opinions led to the conclusion that a such a comparison is not yet possible. This is due to a lack of *comparable* information, due to incomparability of operationalization of testing properties (e.g., test-retest intervals ranging from 1 hour to 1 year) on the one hand, and employment of the questionnaires with incomparable populations on the other hand.

TABLE 3.5 Comparison of characteristics of SIP, NHP, COOP/WONCA, MOS-20/24, SF-36, EuroQol

	SIP	NHP	COOP/WONCA	MOS-20/24	SF-36	EuroQol
Suitable for self-assessment	yes	yes	yes	yes	yes	yes
Number of items	136	38 + 7	6	20 (24)	36	6
Completion time (min.)	20	5	5	5	10	2
Response type	yes/no	yes/no	5 categories with pictogrammes	more than one type (3 - 6 categories)	more than one type (2 - 6 categories)	3 categories + thermometer
Designed for age group...	adults	adults	> 14 years	adults	> 16 years	> 12 years
Designed for ...	varied groups	general population	patients consulting their GP	varied groups	varied groups	varied groups

3.6 Conclusion

A judgement of the contents of the questionnaires revealed that the concept of health status was operationalized somewhat differently in the 6 generic questionnaires for health status assessment currently available for use in the Netherlands. A comparison of their reliability and validity on the basis of literature data was not possible.

It became clear that none of the presently available instruments is superior to all others judged on the basis of 'objective' criteria. This means that none of these instruments is eligible as the standard instrument to be used in medical evaluation research. Even if a 'superior instrument' could be defined, a quick and easy acceptance (i.e., without pressure) to replace the generic instruments currently in use is unlikely. This is due to the investments made by researchers and research groups by using instruments in their previous research.

A general lack of accessible information with respect to the *relative* behaviour of the available instruments in different patient groups was observed. If, for example, a researcher wants to know which of the generic instruments available best suits the purpose of evaluating a new treatment for multiple sclerosis, this information is not easy to find. Maybe the Clearing Houses that have recently been established will be able to fill this information gap to some extent, preferably at a reasonable cost.[23,24]

More 'parallel research' employing two or more instruments with comparable contents in the same patient group is required to provide empirical evidence of the relative value of the instruments. For comparability of questionnaire properties a minimum amount of information should be available in a standard format.

References

1 Bergner M. *Development, testing and use of the Sickness Impact Profile.* In: Walker SR, Rosser R. Quality of Life Assessment: key issues in the 1990s. Dordrecht: Kluwer Academic Publishers, 1993.

2 Luttik A, Jacobs HM, Witte LP de. *De Sickness Impact Profile.* Vakgroep Huisartsgeneeskunde Rijksuniversiteit Utrecht / Instituut voor Revalidatievraagstukken Rijksuniversiteit Limburg, 1987.

3 Melker RA de, Touw-Otten F, Jacobs HM, Luttik A. *De waarde van de 'sickness impact profile' als uitkomstmeting.* Ned Tijdschr Geneeskd 1990;134:946-948.

4 Bruin AF de, Diederiks JPM, Witte LP de, Stevens FCJ, Philipsen H. *The development of a short generic version of the Sickness Impact Profile.* J Clin Epidemiol 1994;47:407-418.

5 Bruin AF de, Diederiks JPM, Witte LP de, Stevens FCJ, Philipen H. *SIP 68 - Handleiding.* Maastricht: Rijksuniversiteit, Vakgroep Medische Sociologie, 1994.

6 McEwen J. *The Nottingham Health Profile.* In: Walker SR, Rosser R. Quality of Life Assess-ment: key issues in the 1990s. Dordrecht: Kluwer Academic Publishers, 1993.

7 Erdman RAM, Passchier J, Kooijman M, Stronks DL. *The Dutch version of the Nottingham Health Profile: investigation of psychometric aspects.* Psychological Reports 1993;72:1027-1035.

8 Essink-Bot ML, Agt HME van, Bonsel GJ. *NHP of SIP: een vergelijkend onderzoek onder chronisch zieken.* T Soc Gezondheidszorg 1992;70:152-159.

9 Weel C van. *Functional status in primary care: COOP/WONCA charts.* Disability and Rehabilitation 1993;15:96-101.

10 Scholten JHG, Weel C van. *Functional status assessment in family practice - the Dartmouth COOP Functional Health Assessment Charts/WONCA.* Lelystad: Meditekst, 1992.

11 Stewart AL, Ware JE, eds. *Measuring functioning and well-being.* London: Duke University Press, 1992.

12 Kempen GIJM. *Het meten van de gezondheidstoestand van ouderen; een toepassing van de Nederlandse versie van de MOS-schaal.* Tijdschrift voor Gerontologie en Geriatrie 1992;23:132-140.

13 Ware JE, Sherbourne CD. *The MOS 36-item Short Form Health Survey (SF-36). I. Conceptual framework and item selection.* Med Care 1992;30:473-481.

14 Bindman AB, Keane D, Lurie N. *Measuring health changes among severely ill patients.*

Med Care 1990;28:1142-1151.

15 Zee K. van der, Sanderman R, Heyink J. *De psychometrische kwaliteiten van de MOS 36-item Short Form Health Survey (SF-36) in een Nederlandse populatie.* T Soc Gezondheidszorg 1993;71:183-191.

16 Zee KI van der, Sanderman R. *Het meten van de algemene gezondheidstoestand met de RAND-36 - een handleiding.* Noordelijk Centrum voor Gezondheidsvraagstukken, Rijksuniversiteit Groningen, 1994.

17 Essink-Bot ML, Stouthard MEA, Bonsel GJ. *Generalizability of valuations on health states collected with the EuroQol questionnaire.* Health Economics 1993;2:237-246.

18 EuroQol Group. *EuroQol: a new facility for the measurement of health-related quality of life.* Health Policy 1990;16:199-208.

19 Deyo RA, Diehr P, Patrick DL. *Reproducibility and responsiveness of health status measures - statistics and strategies for evaluation.* Controlled Clinical Trials 1991;12:142S-158S.

20 Hays RD, Hadorn D. *Responsiveness to change: an aspect of validity, not a separate dimension.* Quality of Life Research 1992;1:73-75.

21 Streiner DL, Norman GR. *Health measurement scales - a practical guide to their develop-ment and use.* Oxford: Oxford Medical Publications, 1989.

22 McDowell I, Martini CJM, Waugh W. *A method for self-assessment of disability before and after hip replacement operations.* BMJ 1978;2:231-246.

23 Erikson P, Scott J. *The On-Line Guide to Quality of Life Assessment (OLGA): resource for selecting quality of life assessments.* In: Walker SR, Rosser RM, eds. Quality of life assessment - key issues for the 1990s. Dordrecht: Kluwer Academic Publishers, 1993.

24 Long AF, Bate L, Sheldon TA. *The outcomes agenda: contribution of the UK clearing house on health outcomes.* Quality in Health Care 1993;2:49-52.

4

NHP or SIP - a comparative study

Accepted for publication in Quality of Life Research as:
Essink-Bot ML, Krabbe PFM, Agt HME van, Bonsel GJ. NHP or SIP - a comparative study in renal insufficiency associated anemia. Reprinted with the permission of the publisher.

4.1 Abstract

In this study we compared the feasibility, internal structure and psychometric characteristics (internal consistency, test-retest reliability, construct validity) of two widely used generic health status measures, i.e., the Nottingham Health Profile (NHP) and the Sickness Impact Profile (SIP) when employed among a sample of patients on renal dialysis (n = 63).
The NHP was found to be more feasible, i.e., shorter and less difficult, than the SIP. The NHP scales showed somewhat higher levels of internal consistency (mean α = .67, range = .39 to .80) than the SIP scales (mean α .65, range = .14 to .82). Test-retest reliability with a 24-hour interval was acceptable for most NHP scales (not available for the SIP in this study). Intercorrelations between the NHP scales were somewhat weaker than those for the SIP, and the expected patterns of scale intercorrelations were largely confirmed. The overall pattern of correlations *between* NHP scales and SIP scales was consistent with expectations, although the correlations were generally rather weak. Correlations between NHP scales and SIP scales and instruments measuring mainly physical functioning (ADL, Karnofsky) were largely as expected. Similarly, correlations between NHP scales and SIP scales and instruments measuring mainly psychological functioning [STAI (anxiety), SDS-Zung (depression)] were also as expected, although here the correlations were weaker for the SIP when compared with the NHP. The Index of Well-being exhibited intra-class correlations >0.3 with one SIP scale and with 5 out of 6 NHP scales. Common factor analysis, yielding a two-factor solution with a physical and a mental factor of equal importance, showed the SIP scales to load more on the physical factor, while the NHP scales loaded more on the mental factor.
The NHP generally performed better than the SIP in terms of feasibility and internal consistency. Physical functioning is emphasized in the SIP, whereas the emphasis of the NHP lies on mental functioning. The analysis also confirms to some extent the intentions of the constructors of NHP and SIP respectively, i.e., the NHP to be a measure of perceived health and the SIP to be a more functional measure.

4.2 Introduction

The assessment of the consequences of disease and treatment on quality of life has gained widespread application. Quality of life in the context of disease and treatment is generally limited to 'health-related quality of life', which is commonly referred to as 'health status'. Health status can be comprehensively operationalized as physical, psychological and social functioning. Examples of applied quantitative health status measurement include the National Health Interview Surveys, research in which the effectiveness of drugs is evaluated, as well as medical technology assessment (MTA) of costly intervention programmes. Data are commonly collected by administering a questionnaire to the subject whose health status is to be measured.

It has become common practice, especially in health status measurement in the context of MTA, to employ a combination of generic instruments with disease and/or domain specific ones. Generic instruments, being comprehensive and non disease specific, allow for the comparison of results among disease stages, and among different diagnostic categories.

Each of the currently available generic instruments has its own strengths and weaknesses. There is, however, little empirical information available on the *relative* performance of these instruments. We hope that the present paper will contribute to the existing knowledge base by addressing an empirical comparison of two generic instruments for measuring health status, i.e., the Nottingham Health Profile (NHP) and the Sickness Impact Profile (SIP).

The specific research questions addressed in this study were:

1. How do the NHP and the SIP compare in terms of feasibility?
2. How do the NHP and the SIP compare in terms of reliability?
3. Is there empirical support for the hypothesized structures of the NHP and the SIP in terms of the health status domains being addressed (i.e., construct validity)?

Quantative analyses of patient data were combined with qualitative research of the questionnaires and literature research. For this purpose we could make use of an existing dataset from a group of patients with renal insufficiency who were treated by renal dialysis. The disease and the intervention have variable consequences for functioning in the physical, psychological and social domains.

4.3 Methods

4.3.1. Instruments

The *Nottingham Health Profile* was developed in the 1970s in the United Kingdom as a measure of perceived health for use in population surveys.[1] The NHP (part 1) consists of 38 dichotomous items which are grouped into six scales, labelled respectively Physical Mobility, Energy, Pain, Sleep, Social Isolation and Emotional Reaction. Each scale ranges from 0 (= optimal) to 100. The ultimate score has a profile format. The Dutch adaptation of the NHP used in the current study has been previously tested in several patient populations. Some NHP items are shown in Table 4.3.1.

TABLE 4.3.1 Examples of NHP items (Hunt 1986)[21]

I have trouble getting up and down stairs or steps (Physical Mobility)
I'm tired all the time (Energy)
I'm in pain when I walk (Pain)
I'm waking up in the early hours of the morning (Sleep)
The days seem to drag (Emotional Reactions)
I feel that I am a burden to people (Social Isolation)

The *Sickness Impact Profile* was developed in the US between 1972 and 1981 as an instrument to assess the consequences of disease and treatment in functional terms. The 136 items are grouped into twelve scales: sleep and rest, eating, work, home management, recreation and pastime, ambulation, mobility, body care and movement (scores of the latter three may be combined as a physical subscore), social interaction, alertness behavior, emotional behavior, communication (scores of the latter four may be combined as a psychosocial subscore). Apart from a 12-dimensional profile score and the physical and psychosocial subscores, the SIP provides the opportunity to compute a total score. Each score ranges from 0 (= optimal) to 100. In the self-assessment version of the SIP the respondent is requested to tick the statements that apply to him/her in relation to his/her health. The SIP was adapted into Dutch by researchers of the Utrecht Institute for General Practice.[2,3] Some examples of SIP items are shown in Table 4.3.2.

Data on 5 additional instruments were used in the investigation of the construct validity of the NHP and the SIP.
The *State-Trait Anxiety Inventory* (STAI) is an American 20-item questionnaire, of which a validated and normed Dutch version is available (ZBV).[4,5] We used the 'state'-part, which measures situational anxiety.[6] The total score ranges from 20 (= no anxiety) to 80.
The *Self-rating Depression Scale* (SDS-Zung) is an American 20-item instrument for measuring depression, with a total score ranging from 25 (= no depressive state) to 100.[7] We used the Dutch version as recommended by the Dutch Psychiatric Society (Vereniging voor Psychiatrie).[8,9]
The *Karnofsky Performance Scale* (or Index) was developed by Karnofsky in 1948 to enable quantification of 'objective' quality of life aspects in the evaluation of drugs against cancer.[10] In the original index, the levels are labelled with figures 0 (= dead), 10, ..., 100 (= optimal). We translated the original US version and adapted it to make it suitable for self-assessment.

TABLE 4.3.2 Examples of SIP items (Bergner, 1981)[22]
I sleep or nap during the day (Sleep and Rest, SR)
I am eating no food at all, nutrition is taken through tubes or intravenous fluids (Eating, E)
I often act irritable toward my work associates (Work, W)
I am not doing any of the maintenance or repair work around the house that I usually do (Home management, HM)
I am going out for entertainment less (Recreation and pastimes, RP)
I walk shorter distances or stop to rest often (Ambulation, A)
I stay away from home only for brief period of time (Mobility, M)
I am very clumsy in body movements (Body care and movement, BCM)
I isolate myself as much as I can from the rest of the family (Social interaction, SI)
I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things (Alertness behavior, AB)
I act irritable and impatient with myself, for example, talk badly about myself, swear at myself, blame myself for things that happen (Emotional behavior, EB)
I am having trouble writing or typing (Communication, C)

Independency with respect to *Activities of Daily Life* (ADL) was assessed by a Dutch instrument asking whether the respondent is able to conduct 9 activities independently, and if so, at which effort. The 9 activities are listed as: getting in and out of bed, going to the lavatory, washing oneself, dressing, eating and drinking, taking a short walk, taking steps, cycling, shopping and cooking. The summary score ranges from 1 to 10 (= completely ADL independent).[11]

The *Index of Well-Being* (IWB) is a measure for subjective well-being which was developed for American population surveys with a score range from 2.1 to 14.7 (= optimal well-being). It was adapted into Dutch.[12]

4.3.2 Patients

We used patients' data from a study to evaluate the effectiveness of erythropietin (EPO) in the treatment of renal insufficiency associated anemia. Questionnaire administration took place around a dialysis session. Before a dialysis session the assessment included completion of a comprehensive questionnaire, which included the NHP but excluded the SIP. The SIP was completed 24 hours later. This second questionnaire also included the NHP in a sample of the patients to investigate test-retest reliability. We did not collect SIP test-retest data because it was considered too burdensome for the patients.

The optimal test-retest interval has to be short enough to preclude a change in health status on the one hand, but long enough to eliminate recollection effects. A change in health status is imaginable between the assessments mentioned above, just preceding dialysis and 24 hours afterwards, respectively. When asked, patients and clinicians generally judged this change as insignificant in relation to the overall health status effects associated with terminal renal insufficiency. Recollection effects can probably be ignored, especially because the NHP was part of a comprehensive questionnaire at the test-assessment.

In the present analyses data were available from 63 patients. Although the study included administration of questionnaires in a longitudinal design, we used data from one administration per patient to prevent introduction of artificial dependence in the data. We had 13 assessments preceding EPO treatment and 50 assessments 1 to 36 weeks after the start of EPO treatment. The mean age of the respondents was 54 years (s.d. 16 years, range 21 - 78 years), 35 (56%) of them were men.

4.3.3 *Statistics*

Features of score distribution. Mean scores, standard deviations, and the percentages of respondents with the best possible score and the worst possible score, respectively, were computed.

The *internal consistency* was determined with Cronbach's α-coefficient. An α-coefficient of 0.70 or higher was considered as sufficient for the purpose of group comparisons.[13]

Test-retest reliability was assessed with the intraclass correlation coefficient (ICC). The ICC is a statistic comparable with the conventional Pearson's correlation coefficient, with level effects between variables being taken additionally into consideration.[14,15] Exact standards for the required magnitude of the reliability coefficient (is the instrument reliable enough?) are difficult to give. A test used for individual judgement should be more reliable than one used for group decisions. Whether a level of test-retest reliability of a test is acceptable for comparisons among groups depends on the size of the group under study: a sample of 1000 can tolerate a less reliable instrument than a sample of 10.[16]

The *internal structure* of the NHP and the SIP was examined with the use of correlation techniques. Matrices of intraclass correlation coefficients (ICCs) between the NHP scales and between the SIP scales, respectively, were computed. For each questionnaire scale, the square root of the mean of the squared ICC between that scale and each of the other scales was computed to summarize the correlation matrix. This statistic was used instead of simply averaging ICCs, in order to retain the interpretation of the squared ICC as the amount of variance shared.

Three approaches were taken to investigate the *construct validity* of the NHP and the SIP. Firstly, the pattern of ICCs between the scales of the NHP and the SIP were examined. It was hypothesized that those scales that are conceptually related would be strongly correlated, while those scales with less in common would exhibit weaker correlations. Secondly, correlation patterns as observed between the scales of the NHP and the SIP and the STAI, the SDS-Zung, the ADL, the Karnofsky and the IWB were compared with a priori hypotheses with respect to these correlation patterns. Thirdly, common factor analysis with varimax rotation was employed to examine the relationships among the elements of the two health status measures and the five additional instruments.

4.4 Results

4.4.1 *Feasibility*

The meaning of the feasibility of questionnaires is not uniformly defined. Some aspects of the NHP and the SIP, considered by the authors to be determinants of 'feasibility' are addressed below.

Item content: the NHP items refer mainly to 'generic' physical and mental *actions*, including for example walking, standing, bending, sleeping, making contact with others, so that the items are applicable to a broad range of age groups, persons in different phases of their lives, and to both sexes. The SIP-items refer to a larger extent to *activities*, including for example tying shoe laces, performing household tasks, lying in bed, performing paid work, visiting friends, caring for children.

Instructions: the SIP instructs respondents to tick the statements which apply to him/her in relation to his/her health. The NHP asks respondents to tick 'yes' if they have the problem stated in each item.The addition of 'in relation to his/her health' contributes to the complexity of the SIP.

Routing: routing refers to conditional questions following responses to preceding questions. There is no routing in the NHP; all respondents must answer all questions. The inclusion of routing in the SIP for Work items adds to the complexity of the instrument and our data did in fact confirm that the respondents were confused. For example, although only 22 respondents indicated that they performed paid work, the SIP Work-items were answered by 44 respondents. Because of this, the SIP Work dimension was left out of further analyses.

Length: The NHP consists of 38 items. It has been reported that an average of 10 minutes is the completion time for self-assessment. The respondents in the present study needed on average 8 minutes (s.d. 3'). The SIP consists of 136 items, with reports of completion time ranging from 20 to 30 minutes.

Complexity: the reading burden may be indicated by the number of words per item. The NHP-DA consists on average of 8.5 (s.d. 3.9) words per item, the SIP of 11.7 (s.d. 6.3). The SIP contains 16 questions comprising more than 20 words, compared with the NHP where this does not occur.

4.4.2 *Features of score distribution*

Mean scores, standard deviations, and the percentages of the respondents with the maximum possible score and the minimum possible score, respectively, for each instrument are shown in Table 4.4.1. The distributions of the scores of the SIP were even more skewed in the direction of good functioning than those of the NHP.

4.4.3 *Internal consistency and test-retest reliability*

The internal consistency coefficients (Cronbach's α) for NHP and SIP scales respectiveley are shown in Table 4.4.1.

The scales of the NHP yielded somewhat higher internal consistency estimates (mean α = .67; range = 0.39 to 0.80) than those of the SIP (mean α = .65; range = 0.14 to 0.82). The α-coefficients for 3 of the NHP scales [Social Isolation (.39), Sleep (.66) and Energy (.69)] and for 5 of the SIP scales [Sleep and rest (.48), Emotional behavior (.62), Home management (.68), Recreation and pastimes (.66), Eating (0.14!)] fell well below the 0.70 standard recommended for group comparisons.

TABLE 4.4.1 Features of score distribution of internal consistency (Cronbach's α) and 24-hours test-retest reliability (ICC) of NHP and SIP scales; score distributions of STAI, SDS-Zung, ADL, IWB and Karnofsky. Renal dialysis patients, n = 63

	mean	s.d.	% max*	% min**	α	test-retest
NHP (score 0-100)						
Physical Mobility (8)***	26.3	24.8	29	0	.80	.80
Energy (3)	33.0	35.8	43	13	.69	.62
Pain (8)	13.3	20.6	46	0	.76	.73
Sleep (5)	38.6	34.9	24	10	.66	.75
Emotional Reactions (9)	17.6	21.8	38	2	.74	.55
Social Isolation (5)	12.9	19.7	60	0	.39	.57
SIP (score 0-100)						
Sleep & Rest (7)	16.8	17.1	27	0	.48	--
Emotional Behavior (9)	6.5	11.0	67	0	.62	--
Bodycare and Movement	6.7	9.9	38	0	.81	--
Home management (10)	21.7	20.5	21	0	.68	--
Mobility (10)	12.7	14.1	46	0	.70	--
Social Interaction (20)	9.3	9.7	25	0	.75	--
Ambulation (12)	15.4	14.7	29	0	.73	--
Alertness Behavior (10)	11.8	18.5	57	0	.82	--
Communication (9)	6.0	12.6	71	0	.77	--
Recreation & Pastimes (8)	29.5	22.8	16	0	.66	--
Eating (9)	9.4	5.4	13	0	.14	--
SIP total score	12.2	9.5	16	0	.95	--
SIP physical score	9.8	10.6	19	0	.89	--
SIP psychosocial score	8.6	10.2	0	0	.90	--
ADL (score 10-1)	8.8	1.4	44	0	--	--
STAI (score 20-80)	38.6	11.3	3	0	--	--
SDS-Zung (score 25-100)	40.1	8.2	1	0	--	--
Karnofsky (score 100-0)	72.2	16.4	11	0	--	--
IWB (score 14.7-2.1)	10.4	3.2	0	0	--	--

* % max=percentage of respondents with best possible score (ceiling); ** % min=percentage of respondents with worst possible score (floor); *** number of items

Nineteen SIP-items showed zero variance, which was explainable because they addressed very serious impairment of functioning.

Test-retest reliability estimates (ICCs) for the NHP scales are also shown in Table 4.4.1. The precautions mentioned in the Patients-section are to be borne in mind when interpreting these figures. As could be expected from the item content, test-retest reliability was highest for Physical Mobility(NHP). Test-retest reliability was rather low for Social Isolation (NHP) and Emotional Reaction(NHP).

4.4.4 *Structure*

The ICCs for the NHP scales and the SIP scales, respectively, are summarized in Table 4.4.2 (complete data shown in Appendix 1). In general, the NHP scales were somewhat less highly intercorrelated than were the SIP scales. As was expected, high ICCs were observed between Social Isolation(NHP) and Emotional Reaction(NHP). The SIP scales grouped in the Physical subscore (Bodycare & Movement, Mobility, Ambulation) showed high intercorrelations. A similar pattern was observed for the SIP scales grouped into the psychosocial subscore (Social Interaction, Alertness Behavior, Emotional Behavior, Communication). Eating(SIP) correlated low with the other SIP scales.

4.4.5 *Construct validity*

Firstly, the matrix of ICCs between NHP scales and SIP scales is presented in the Appendix. We expected higher correlations between 'physical' dimensions (Physical Mobility(NHP), Bodycare and Movement(SIP), Mobility(SIP), Ambulation(SIP)) and between 'psychosocial' dimensions (Social Isolation(NHP), Emotional Reaction(NHP), Emotional Behavior(SIP), Social Interaction(SIP), Alertness Behavior(SIP), and Communication (SIP)); and weaker correlations between physical and psychosocial dimensions.

The correlations observed between the NHP and SIP scales were generally rather low. There were some deviations from the expected patterns; for example, low ICCs between Social Isolation(NHP) and Emotional Behavior(SIP), between Social Interactions (NHP) and Communication(SIP), between Emotional Reaction(NHP) and Communication (SIP). The latter two observations are understandable as the items of Communication (SIP) are of a rather physical nature (e.g., difficulties in speaking).

Secondly, correlation patterns as observed between the scales of NHP and SIP and 5 instruments with proved validity (STAI, SDS-Zung, ADL, Karnofsky, IWB) were compared to a priori hypotheses with respect to these correlation patterns. For example, we expected the highest correlations with ADL and Karnofsky for Physical Mobility(NHP) and for Pain(NHP), and we expected the highest correlations with STAI and SDS-Zung for Social Isolation(NHP) and Emotional Reaction(NHP). We similarly expected the highest correlations with STAI and SDS-Zung for the components of the psychosocial subscore of the SIP (Social Interaction, Alertness behavior, Emotional behavior and Communication), and the highest correlations with ADL and Karnofsky for the components of the physical subscore of the SIP (Bodycare and movement, Mobility, Ambulation).

The association patterns observed between the NHP and the SIP, respectively, and the other five instruments were largely as expected (see Table 4.4.3). Exceptions were Communication(SIP) which correlated weakly with STAI and SDS-Zung, understandable in view of the from the reasoning described above, and Social interactions(NHP) which also correlated weakly with STAI. The IWB (as a measure for experienced well-being) showed the highest correlations (ICC >.3) with Recreation & pastimes $_{(SIP)}$, Household management$_{(SIP)}$, and all NHP dimensions except Pain$_{(NHP)}$.

TABLE 4.4.2 Internal structure of NHP and SIP: summary* of ICCs for each scale with the other scales of NHP and SIP respectively (renal dialysis patients, n=63)

NHP	Phys. Mob	Pain	Energy	Sleep	Soc	Emot						Total
	.41	.38	.33	.32	.35	.43						.37
SIP	SR	EB	BCM	HM	M	SI	A	AB	C	RP	E	Total
	.39	.39	.43	.43	.48	.45	.43	.39	.38	.32	.22	.40

* For example: the figure of .41 for NHP Physical Mobility represents the square root of $(((.49)^2+(.48)^2+(.40)^2+(.31)^2+(.33)^2)/5)$ (Appendix 1)

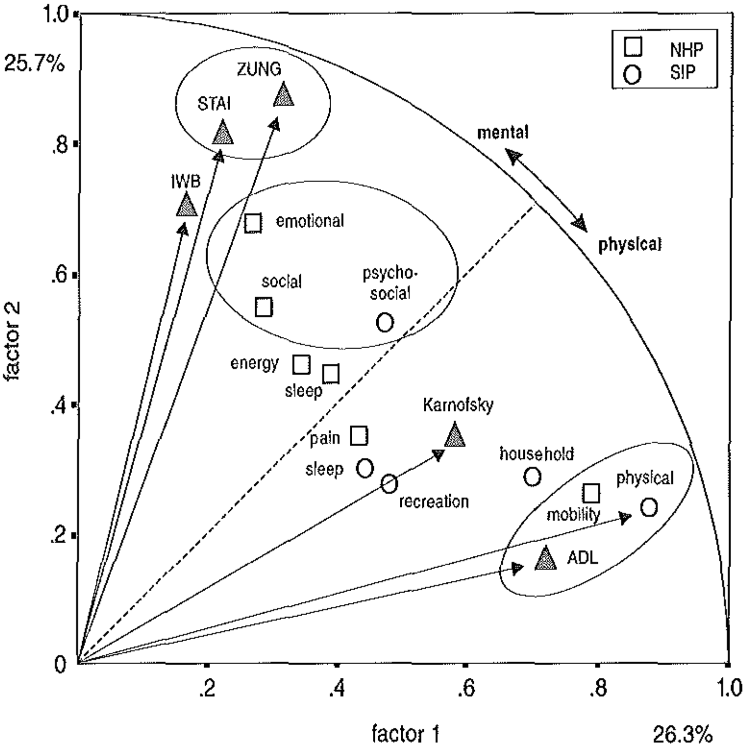
Common factor analysis with varimax rotation of the combined data of NHP (6 scales), SIP (physical subscore, psychosocial subscore, Sleep & rest, Recreation & pastimes, Household management), ADL, Karnofsky, STAI, SDS-Zung and IWB yielded two factors with eigenvalues >1.0; see Figure 4.4. The first factor explained 26.3% of common variance and was interpreted as a physical dimension, the second factor explained 25.7% of common variance and was interpreted as a mental dimension. Scales with high loadings on the physical factor were the Physical subscore of the SIP; Physical Mobility$_{(NHP)}$; ADL; Household management$_{(SIP)}$; and Karnofsky. Scales with high loadings on the mental factor were SDS-Zung; STAI; IWB; Emotional reaction$_{(NHP)}$; Social Isolation$_{(NHP)}$; and the SIP psychosocial subscore. The physical scales of NHP and SIP (Physical Mobility$_{(NHP)}$ and the physical subscore$_{(SIP)}$) are closer to each other in Figure 4.4 than the mental scales (Emotional reaction $_{(NHP)}$, Social Isolation$_{(NHP)}$, psychosocial subscore$_{(SIP)}$). This means that there is more similarity between NHP and SIP in the physical domain than in the mental domain. The IWB loaded very high on the second factor, indicating that well-being as measured with the IWB is largely determined by mental factors in this population.

TABLE 4.4.3 Correlation (ICCs) of NHP and SIP scales, respectively, with STAI, SDS, ADL Karnofsky and IWB (renal dialysis patients, n = 63)

	ADL*	Karnofsky*	STAI*	SDS-Zung*	IWB*
NHP					
Physical mobility	.58	.55	.35	.37	.37
Pain	.41	.32	.25	.35	.23
Energy	.20	.34	.48	.28	.36
Sleep	.25	.30	.32	.24	.39
Social Isolation	.32	.22	.28	.46	.34
Emotional reaction	.27	.35	.48	.48	.37
SIP					
Sleep and rest	.25	.27	.31	.35	.21
Emotional behavior	.16	.15	.22	.27	.14
Bodycare & movement	.55	.20	.13	.28	.12
Home management	.42	.40	.34	.37	.32
Mobility	.57	.34	.22	.29	.10
Social interaction	.20	.16	.23	.34	.17
Ambulation	.51	.32	.20	.37	.20
Alertness behavior	.19	.27	.30	.41	.24
Communication	.18	.18	.08	.17	.04
Recreation & pastimes	.16	.35	.35	.22	.33
Eating	.05	.07	.04	.06	.02

* Rescaled to a 0-100 scale (0 = optimal score) in accordance with NHP and SIP scales.

FIGURE 4.4 Factor analysis with varimax rotation of NHP, SIP, ADL, Karnofsky, STAI, SDS-Zung and IWB (renal dialysis patients, n=63)



4.5 Conclusion and discussion

In this study we have compared the feasibility, structure and psychometric characteristics of 2 well-known generic health status measures - the NHP and the SIP - when employed in a group of renal dialysis patients. The results are summarized in Table 4.5.

TABLE 4.5 Summary of the empirical comparison of NHP and SIP

	NHP	SIP
Feasibility	generally better	
Internal consistency	acceptable for 5 out of 6 scales	acceptable for 8 out of 11 scales
Test-retest reliability	acceptable	*not available*
Structure	confirmed	confirmed
Construct validity	more emphasis on mental health, perceived health	more emphasis on physical health, functional health

The NHP can be considered to be generally more feasible than the SIP. The NHP is shorter and less difficult. The observed difference in item contents (relating to actions in the NHP, to activities in the SIP) might cause the SIP to be less universally applicable and more culture-bound than the NHP. For example, the Work items of the SIP have often been observed to be omitted from the questionnaire in elderly populations. It is interesting to note that Part 2 of the NHP, which was not used in the empirical part of our study and is thus not addressed in this paper, refers to activities as well.

The results for internal consistency were better for the NHP than for the SIP. The internal consistency is (almost) acceptable for 5 out of 6 NHP scales, and for 8 out of 11 SIP scales. Published data on internal consistency of the NHP scales for the UK version appeared to be unavailable. The study by Erdman et al among 276 Dutch general practice patients showed a mean α of 0.78, all αs 0.70 or higher.[17] The lower internal consistency estimates in our study especially for the Social Isolation Scale (0.39), may be due to the different nature of the study population. It supports the fact that psychometric characteristics are population-specific.

Internal consistency estimates for 10 out of 12 US SIP scales are available for a stratified sample of members of a US prepaid group practice [n = 495; mean α = .61, range = 0.29 (Eating) to 0.82 (Social interaction); 8 out of 10 αs below 0.70] and a group of 168 noncognitively impaired nursing home patients [mean α = 0.72, range = 0.60 (Eating, Sleep and rest) to 0.84 (Body care and movement); 3 out of 10 αs below 0.70].[18] These results and the results of the present study are indicative of a borderline acceptable level of internal consistency of several SIP scales. Internal consistency estimates for the SIP as a whole (136 items) exceed .90 for the US, the Swedish, the Spanish and the Dutch version, but this is partly attributable to the large number of items.[19]

With respect to test-retest reliability, results (4-week intervals) for the UK NHP among 58 arthrosis patients were in the range of 0.77 to 0.85 (Spearman rank correlation coefficients)[20,21] and among 93 patients with peripheral vascular disease in the range of 0.75 to 0.88[1]. Test-retest reliability of the Dutch NHP in cardiac patients showed Spearman rank correlations of 0.69 - 0.84.[17] The somewhat lower test-retest reliability estimates in the present study may be partly attributed to the fact that it is not quite sure that patients health status remained unchanged between the two assessments: preceding dialysis and 24 hours later. For the US SIP, 24 hours test-retest reliability coefficient was 0.92 for the total score over 136 items.[22]

Examination of the inter-scale correlations for the NHP and the SIP showed these correlations to be of moderate magnitude, suggesting little redundancy of information generated by the scales of the instruments. For the Dutch NHP, these results replicate the findings of Erdman et al.[17]

The ICCs observed *between* NHP scales and SIP scales were rather low, suggesting that the NHP and SIP to some extent measure different aspects of health status. ICCs observed between the NHP scales and the SIP scales, respectively, and instruments indicating mainly physical functioning (ADL, Karnofsky) and mainly psychological functioning (STAI, SDS-Zung) were largely as expected. However, the psychosocial scales of the SIP correlated more weakly with STAI and SDS-Zung than the psychosocial scales of the NHP. The IWB exhibited ICCs >0.3 with one SIP scale and with 5 out of 6 NHP scales. Factor analysis yielded a two-factor solution with a physical and a mental factor of equal importance and showed the SIP scales to load more on the physical factor (with the psychosocial subscore as the only exception). A similar result was obtained by Bruin et al, who performed principal components analysis on 835 SIPs completed by subjects from different diagnostic categories.[23]

The NHP scales loaded more on the mental factor (exceptions: Physical Mobility, Pain). This may be interpreted as the SIP emphasizing physical functioning, whereas the NHP emphasizes mental functioning. The analysis also confirms to some extent the intentions of the constructors of the NHP and the SIP respectively, i.e., that the NHP was intended to be a measure of perceived health while the SIP was intended to be a more functional measure.

The results of the present study add to the developing body of knowledge with respect to performance characteristics of Dutch adaptations of the NHP and the SIP. A cross-culturally adapted health status measure is essentially a new instrument, and investigation of its characteristics is required.[16] Cross-cultural adaptation of health status measures requires more than 'conceptually equivalent' translation, because of expected cultural differences with respect to health beliefs and response to questionnaires. This is required even among residents of industrialized societies. Jacobs showed that the US item weights for the SIP items can be validly applied for Dutch SIP data.[24] The French NHP item weights showed some differences if compared with the British ones.[25]

The NHP generally performed better than the SIP in this study. This does *not* imply that the NHP is generally to be preferred to the SIP in medical evaluation research. Firstly, responsiveness to change over time was not a subject of comparison in the present study. Secondly, performance characteristics of generic instruments for health status are probably population specific. For an instrument to perform well it must do so in terms of feasibility, internal consistency, test-retest reliability, and validity including responsiveness to change over time. An instrument which performs well according to the aforementioned criteria in a population of elderly, rather seriously ill patients with renal insufficiency will not necessarily perform equally well when employed for example among young patients with lung problems. The possibility that an instrument performs equally well in all types of patient groups with varying degrees of illness can be seriously doubted. The case might eventually be that NHP and SIP are each superior in different groups.

Notes

1. An exception to the broad applicability of the NHP was observed when the NHP was employed in another study among patients with spinal cord injury. As these patients were not able to walk at all, most of the items belonging to the dimensions Physical Mobility and Pain were 'not applicable' for them.

2. Eating$_{\text{(SIP)}}$ was left out of the ultimate factor analysis that is presented here, because it was so different from the other variables (see low correlations with the other variables) that it emerged as a separate 'factor' and interfered too much with the factor analysis.

Acknowledgement

The authors wish to thank the respondents who participated in this study, Hans Severens MSc and Professor Eddy KA van Doorslaer PhD for their respective contributions to the present study, as well as the Advisory Board for the Health Research Promotion Programme (Adviesgroep SGO) for supporting financially the Research Programme 'Standardization in Medical Technology Assessment'.

APPENDIX 1 ICCs of NHP scales and SIP scales (n=63).

	PHYSMOB	PAIN	ENERGY	SLEEP	SOCIAL	EMOTION	SR	EB	BCM	HM	M	SI	A	AB	C	RP
PAIN	.49															
ENERGY	.48	.25														
SLEEP	.40	.23	.33													
SOCIAL	.31	.38	.12	.24												
EMOTION	.33	.48	.38	.38	.56											
SR	.27	.20	.09	.22	.23	.23										
EB	.10	.31	.11	.09	.22	.38	.35									
BCM	.32	.42	.10	.11	.34	.24	.29	.50								
HM	.56	.28	.29	.26	.35	.35	.46	.24	.34							
M	.40	.31	.18	.15	.32	.33	.51	.47	.60	.61						
SI	.17	.14	.14	.11	.38	.31	.51	.56	.48	.38	.50					
A	.43	.36	.23	.13	.21	.27	.40	.38	.54	.58	.61	.36				
AB	.25	.34	.19	.11	.59	.41	.28	.42	.41	.43	.41	.50	.38			
C	.16	.15	.05	.05	.27	.25	.22	.42	.51	.30	.42	.44	.26	.57		
RP	.37	.08	.33	.23	.17	.26	.46	.20	.14	.54	.34	.26	.42	.29	.17	
E	.02	.05	.02	.02	.02	.07	.21	.23	.21	.14	.20	.40	.22	.16	.14	.12

PHYSMOB=NHP Physical Mobility; PAIN=NHP Pain; ENERGY=NHP Energy; SLEEP=NHP Sleep; SOCIAL=NHP Social Isolation; EMOTION=NHP Emotional Reaction; SR=SIP Sleep & rest; EB=SIP Emotional behavior; BCM=SIP Bodycare & movement; HM=SIP Home management; MOB=SIP Mobility; SI=SIP Social interaction; A=SIP Ambulation; AB=SIP Alertness behavior; C=SIP Communication; RP=SIP Recreation & pastimes; E=SIP Eating.

References

1 Hunt SM, McEwen J, McKenna SP. *Measuring health status: a new tool for clinicians and epidemiologists.* J R Coll Gen Pract 1985;35:185-188.

2 Luttik A, Jacobs HM, Witte LP de. *De Sickness Impact Profile.* Vakgroep Huisartsgeneeskunde Rijksuniversiteit Utrecht / Instituut voor Revalidatievraagstukken Rijksuniversiteit Limburg, 1987.

3 Meiker RA de, Touw-Otten F, Jacobs HM, Luttik A. *The value of the 'Sickness Impact Profile' in outcome measurement. (De waarde van de 'Sickness Impact Profile' als uitkomstmeting, in Dutch).* Ned Tijdschr Geneeskd 1990;134:946-948.

4 Ploeg HM van der, Defares PB, Spielberger CD, ed. *Manual of the State-Trait Anxiety Inventory (Handleiding bij de zelfbeoordelingsvragenlijst).* Leiden: Swets & Zeitlinger, 1980.

5 Ploeg HM van der. *Validation of de State-trait Anxiety Inventory (Validatie van de zelfbeoordelingsvragenlijst, in Dutch).* Ned T Psychol 1980;35:243-249.

6 Spielberger CD, Gorschuch RL, Lushene RE (eds). *STAI manual for the state-trait anxiety inventory.* Consulting Psychologuists Press, Palo Alto, California 1970.

7 Zung WWK. *A self-rating depression scale.* Arch Gen Psych 1965;13:63-70.

8 Zitman FG, Griez EJL, Hooijer Chr. *Standardisation of depression assessment questionnaires (Standaardisering depressievragenlijsten; in Dutch).* T Psychiatr 1989;31:114-135.

9 Dijkstra P. *The Zung self-rating depression scale (De zelfbeoordelingsschaal voor depressie van Zung, in Dutch).* In: HM van Praag & HGM Rooymans, Stemming en ontstemming, pp. 98-120. Amsterdam: de Erven Boon, 1974.

10 Karnofsky DA, Abelman WH, Craver LF, Burchenal JH. *The use of nitrogen mustards in the palliative treatment of carcinoma.* Cancer 1948;643-654.

11 Bonsel GJ, Bot ML, Boterblom A, Veer F van 't. *Costs and effects of heart transplantation (De kosten en effecten van harttransplantatie), part 2A, 2B, 2C: Quality of life - documentation, interview, results (Kwaliteit van leven voor en na harttransplantatie - documentatie, interview, resultaten, in Dutch).* Department of Public Health, Erasmus University, Rotterdam 1988.

12 Campbell A, Converse PE, Rodgers WL. *The quality of American life: perceptions, evaluations and satisfactions.* Russell Sage Foundation, New York 1976.

13 Nunnally JC. *Psychometric theory.* McGraw Hill, New York 1978.

14 Dunn G. *Design and analysis of reliability studies.* Oxford: Oxford University Press, 1989.

15 Deyo RA, Diehr P, Patrick DL. *Reproducibility and responsiveness of health status measures.* Controlled Clinical Trials 1991;12:142S.

16 Streiner DL, Norman GR. *Health measurement scales*. Oxford: Oxford Medical Publications, 1989 (ISBN 0 19 2617737).

17 Erdman RAM, Passchier J, Kooijman M, Stronks DL. *The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects*. Psych Reports 1993;72:1027-1035.

18 Rothman ML, Hedrick S, Inui T. *The Sickness Impact Profile as a measure of the health status of noncognitively impaired nursing home residents*. Med Care 1989;27(Suppl 3):S157-S167.

19 Bruin AF de, Witte LP, Stevens FCJ, Diederiks JPM. *The usefullness of the Sickness Impact Profile as a generic fuctional status measure (De bruikbaarheid van de Sickness Impact Profile als generieke maat voor de gezondheidstoestand, in Dutch; English abstract)*. T Soc Gezondheidsz 1992;70:160-170.

20 Hunt SM, McKenna SP, Williams J. *Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis*. J Epidemiol Community Health 1981;35:297-300.

21 Hunt SM, McEwen J, McKenna SP. *Measuring Health status*. London, Croom Helm, 1986.

22 Bergner M, Bobbitt RA, Carter WB, Gilson BS. *The Sickness Impact Profile: development and final revision of a health status measure*. Med Care 1981;19:787-805.

23 Bruin AF de, Diederiks JPM, Witte LP de, Stevens FCJ, Philipsen H. *The development of a short generic version of the Sickness Impact Profile*. J Clin Epidemiol 1994;47:407-418.

24 Jacobs HM, Luttik A, Touw-Otten FWMM, Melker RA de. *The Sickness Impact Profile: validation of the Dutch version. (De 'Sickness Impact Profile': resultaten van een valideringsonderzoek van de Nederlandse versie, in Dutch)*. Ned Tijdschr Geneeskd 1990;134:1950-1952.

25 Bucquet D, Condon S, Ritchie K. *The French version of the Nottingham health profile. A comparison of items weights with those of the source version*. Soc Sci Med 1990;30:829-835.

5

An empirical comparison of 4 generic health status measures: the Nottingham Health Profile, the MOS 36-Item Short-Form Health Survey, the COOP/WONCA Charts, and the EuroQol Instrument

Submitted as:
Essink-Bot ML, Krabbe PFM, Bonsel GJ, Aaronson NK. An empirical comparison of 4 generic health status measures: the Nottingham Health Profile, the MOS 36-Item Short-Form Health Survey, the COOP/WONCA Charts, and the EuroQol instrument.

5.1 Abstract

In this study we compared the feasibility, internal structure and psychometric characteristics (internal consistency, construct validity, 'known groups' validity) of 4 generic health status measures - the NHP, the SF-36, the COOP/WONCA charts and the EuroQol - when employed in a sample of migraine sufferers and a control group (total n=1,011).

In terms of feasibility, the NHP had the lowest missing value rate. The SF-36 exhibited high levels of internal consistency (αs between .76 and .91) as compared with the NHP (.62 - .82). The NHP scales were somewhat less highly intercorrelated than those of the SF-36. The COOP/WONCA items were found to be relatively highly correlated with one another, as were the EuroQol items. The overall patterns of correlations *between* the scales of the NHP and the SF-36, and *between* the COOP/WONCA charts and the EuroQol items, respectively, were consistent with expectations. Two combined factor analyses (i.e., the SF-36 scales, the COOP/WONCA items and the EuroQol items; the SF-36 and NHP scales, and the EuroQol items) resulted in similar solutions, with two higher-order factors being identified - one reflecting physical health, the other mental health. A qualitative comparison of the measures indicated that each addresses these two basic health domains from a somewhat different perspective. The SF-36 was best able to discriminate between groups formed on the basis of self-reported chronic conditions and work disability days, respectively. In general, all four instruments exhibited a good performance profile. However, both instruments with a multi-item structure (i.e., the SF-36 and the NHP) outperformed the COOP/WONCA charts and the EuroQol. Future research is needed to investigate the relative performance of these measures when employed in more seriously ill patient populations, and to extend the comparison to test-retest reliability and responsiveness to change in health over time.

5.2 Introduction

Generic health status assessment was applied first in population surveys measuring the state of health of communities, irrespective of diagnosis. More recently, researchers have recognized the potential value of incorporating generic health status outcomes in the evaluation of medical interventions, in addition to more traditional biologic (e.g., survival) and symptom-oriented measures. Although every disease is associated with specific health effects, these effects are also reflected in patterns of impairment at the broader level of physical, psychological and social functioning captured by the more generic class of health status measures. Because generic outcome measures can be used to evaluate the functional health of individuals without regard to cause[1], they offer the opportunity of comparing levels of functioning across patient populations, and between patient populations and the general population. In this way, rank ordering of diseases according to their relative effects on functioning (i.e., burden of disease[2]) and of health care interventions in terms of their impact on functioning levels (i.e., treatment effectiveness) becomes possible.

A range of generic health status measures are currently available, including the Sickness Impact Profile (SIP), the Nottingham Health Profile (NHP), and the more recently developed COOP/WONCA charts, EuroQol and MOS 36-Item Short Form Health Survey (SF-36).[3,4,5,6,7,8,9] Each of these measures has its particular strengths and weaknesses. The decision to use any one of these measures in a particular survey or clinical trial is often based on diverse scientific and extrascientific considerations, including the nature of the research questions to be addressed, the characteristics of the study population, the tradition of the research group, and intellectual investments made in a given instrument in previous research.

Relatively little attention has been paid to the fact that the performance characteristics of an instrument, including feasibility, reliability and validity, may, to a greater or lesser degree, be population dependent. If, for example, an instrument performs well in a population of seriously ill cardiac patients, this does not guarantee that it will work equally well when employed among patients with low backpain. Given the increasing use of formal health status assessment in medical research, there is a pressing need for empirical data on the *relative* performance of the available generic measures among distinct patient populations.

In an effort to contribute to this process, we conducted a study of the health status of migraine sufferers and a matched control group in which a head-to-head comparison was made between the NHP, the SF-36, the COOP/WONCA charts and the EuroQol instrument. Although the study had primarily a substantive research focus, these 4 generic health status instruments were purposively included in the research design to enable these comparisons to be made.

The specific research questions addressed in this study were:

1. How do these 4 instruments compare in terms of feasibility and reliability?
2. Is there empirical support for the hypothesized structure of the questionnaires in terms of health status domains being addressed (i.e., construct validity)?
3. How do the instruments compare in their ability to discriminate between groups known to differ on other indicators of health (e.g., presence of chronic health conditions, disability days) ('known groups' validity)?[10]

5.3 Methods

5.3.1 *Study sample and data collection procedures*

Migraine sufferers and a matched control group were surveyed to assess the societal impact (costs and health status effects) of migraine in the Netherlands. Details of the study design and substantive results are described elsewhere.[11,12] The following provides a brief description of the sampling strategy and data collection procedures.

To establish the prevalence of migraine in the Netherlands, face-to-face interviews were conducted with a representative sample of the Dutch general population (n=10,480). Subjects were included as migraine patients if they met the International Headache Society criteria[13] *and* had experienced at least one attack of migraine during the 12 months prior to the interview. 992 migraine sufferers met these criteria (1-year prevalence = 9.5%). Of these 992 cases, 846 (85%) expressed an initial willingness to take part in a subsequent study investigating the impact of migraine on health status and direct/indirect costs.

The control group was selected from among those subjects in the original prevalence survey who did not meet the criteria for migraine. Frequency matching was used to generate a control group reflecting the age (in 5-year intervals), gender and employment status characteristics of the migraine sample.

Questionnaires were mailed in June, 1993, followed by reminders after 2 and 5 weeks. Half of the addressees received a packet containing the NHP, the SF-36 and the EuroQol. For the other half of the sample, the NHP was replaced by the COOP/ WONCA charts. This was done to reduce the total respondent burden (i.e., a total of 3 rather than 4 questionnaires were administered per respondent). The sequencing of the questionnaires was varied systematically in order to avoid an ordering effect.

The useable response rate was 58% (n=436) in the migraine group and 71% (n=575) in the control group. Non-response analyses failed to reveal any statistically significant differences between addressees and respondents with regard to age, gender, social class or degree of urbanization in either the migraine group or the control group.

The number of questionnaires available from the migraine group and the control group, respectively, was: the SF-36 = 436 and 575; the EuroQol = 436 and 575; the COOP/ WONCA charts = 210 and 286; and the NHP = 226 and 289.

5.3.2 *Health Status Measures*

The *Nottingham Health Profile* was developed in the 1970s in the United Kingdom as a measure of perceived health for use in population surveys.[4] The NHP consists of 38 dichotomous items which are grouped into the scales described in Table 5.4.1. Each scale ranges from 100 to 0 (= optimal). The Dutch adaptation of the NHP used in the current study has been previously tested in several patient populations.[14,15]

The *MOS 36-Item Short-Form Health Survey*, developed in the United States, is derived from the larger battery of health status instruments employed in the Medical Outcomes Study.[8,9,10,16] It consists of 36 items, organized into 8 scales (see Table 5.4.1). The number of response choices per item ranges from 2 to 6. The SF-36 yields an 8-dimensional profile, with each scale having a range from 0 to 100 (=optimal). The Dutch version of the SF-36 employed in the current study was developed as a part of the International Quality of Life Assessment (IQOLA) Project, whose objective is to translate, validate and norm the SF-36 in a wide range of languages and cultural settings.[17]

The *COOP/WONCA charts* were developed to assess the functional status of patients in primary care settings.[5] Subjects are requested to score their functioning on each of the 6 items described in Table 5.4.1 during the 2 weeks prior to assessment on 5-point scales (1=optimal). The levels on the scales are illustrated with pictograms. The standard Dutch version of the revised charts was used.[18]

The *EuroQol instrument* was developed by the international EuroQol Group as a standardized generic measure for description of health status, with the additional possibility of converting the descriptive data into values for economic (cost-effectiveness) analysis by linking patients' health state descriptions to empirical valuations of health states obtained from the general population.[6] The standard Dutch 5D-version of the EuroQol was used.[7] The descriptive part of the instruments consists of 5 items (see Table 5.4.1), each following the general form: 1 = no problems, 2 = some problems, 3 = extreme problems. The 6th item is a global health evaluation using a visual analogue scale ranging from 0 (worst imaginable health state) to 100 (best imaginable health state). Only data from the first 5 items are included in the current analyses.

5.3.3 *Additional variables*

A standard set of sociodemographic questions were asked to obtain information on age, sex, education and employment status.

Comorbidity was assessed by the standard list of chronic conditions of the Central Bureau for Statistics (CBS). This list enumerates 28 conditions in lay terms (e.g., 'asthma, chronic bronchitis or COPD', 'diabetes'). For each chronic condition, respondents were asked to report if they currently had the condition, or if they had had it in the previous year.

5.3.4 *Analysis Plan*

Qualitative analysis of questionnaire content

A qualitative comparison was carried out of the item-content of the multi-item scales of the NHP and the SF-36, and of the individual items of the COOP/WONCA and the EuroQol. Scales/items were considered 'comparable' if their content was judged to refer to the same general health domain.

Quantitative analyses

All of the following analyses were performed for the migraine group and the control group separately, as well as for the pooled data. The results of the analyses based on the pooled data will be presented, except in those cases where the separate analyses yielded significant differences.

Feasibility. The number of missing cases per item was employed as an empirical indicator of feasibility. Missing values were defined as those cases where no answer was provided, and those where multiple responses were given when only one was required. As the number of respondents with complete records was large enough for further analyses, we did not impute constructed values for missing values. For the purpose of comparability, an index was constructed accounting for the number of respondents and the number of items per questionnaire.

Features of score distribution. Mean scores, standard deviations, and the percentages of respondents with the maximum possible score and the minimum possible score, respectively, were computed per scale (NHP, SF-36) or item (COOP/WONCA, EuroQol), respectively.

Reliability. The internal consistency of the NHP and SF-36 multi-item scales was determined with Cronbach's α-coefficient.[19] An α-coefficient of 0.70 or higher was considered as sufficient for the purpose of group comparisons.[19,20] Internal consistency estimates could not be calculated for the COOP/WONCA charts or the EuroQol, as these instruments consist of 1 item with an ordered response choice per 'scale'. Due to the cross-sectional nature of the study, data on test-retest reliability were not available.

Internal scale structure. The internal structure of the 4 instruments was examined with the use of correlation techniques. For the NHP and the SF-36 scales, intraclass correlation coefficients (ICCs) were employed. The ICC is a statistic comparable with the conventional Pearson's correlation coefficient, with level effects between variables being taken into consideration.[21,22] For each questionnaire scale, the square root of the mean of the squared ICCs between that scale and each of the other scales was computed to summarize the ICC-matrix. This statistic was used instead of averaging ICCs, in order to retain the interpretation of the squared ICC as the amount of variance shared. ICCs are not appropriate for the ordinal EuroQol and COOP/WONCA data. As an alternative, polychoric correlation coefficients (PCCs) were used. The PCC has certain statistical advantages over alternative indicators such as the Spearman rank correlation coefficient. First, the PCC provides a reliable estimate of the correlation between ordinal variables even when the number of categories is limited. Second, the PCC does not appear to be sensitive to the shape of the marginal distributions. Finally, the PCC uses the attractive premise of a continuous bivariate normal distribution underlying the categories.[23,24] The PCC correlation matrices were summarized in a manner similar to the ICC matrices.

Construct validity. Two approaches were taken to examining the construct validity of the 4 health status instruments. First, the pattern of correlations between the scales of the NHP and the SF-36 (ICCs), and between the items of the COOP/WONCA and the EuroQol (PCCs) were examined. It was hypothesized that those scales/items that are conceptually related would be relatively strongly correlated, while those scales/items with less in common would exhibit weaker correlations.

Second, common factor analysis with varimax rotation was employed to examine the relationships among the elements of the 4 health status measures, and to look for possible higher order factors. Because any given respondent completed only 3 of the 4 instruments (see section 5.3.1), 2 factor analyses were performed: (1) with data from the NHP (scales), the SF-36 (scales), and the EuroQol (items); and (2) with data from the SF-36 (scales), COOP/WONCA (items) and EuroQol (items) (see Note 1).

'Known groups' validity. A series of statistical tests was carried out to evaluate the ability of the 4 health status measures to discriminate between subgroups of respondents known to differ on several relevant variables. For these group comparisons, the Mann-Whitney U test was employed due to the non-normal distribution of the data. The grouping variables included: (1) the number of self-reported chronic conditions (≤ 1 versus > 1; n=1,011); and (2) for the respondents with paid employment (n=461), the number of days absent from work due to illness in the 2 weeks prior to assessment (0 versus ≥ 0.5 days).

Given a large enough sample size, statistical significance can be somewhat misleading. That is, relatively small mean group differences may reach conventional levels of statistical significance without representing meaningful differences in functioning. For this reason, an effect size estimation was calculated which relates the difference in mean scores to the dispersion of the scores. The formula employed to calculate the effect size (*d*) was: Mean(a) - Mean(b) / standard deviation (see Note 2). Following Cohen's suggested guidelines, $d = 0.2$ was taken to indicate a small effect size, $d = 0.5$ a moderate effect size, and $d = 0.8$ a large effect size.[25]

5.4 Results

The respondents in the migraine group and in the control group were comparable in terms of gender distribution (84% versus 80% female), mean age (40 versus 41 years), employment status (47% versus 44% in paid employment), and educational level (38% versus 38% with an intermediate educational level; 28% versus 31% with a higher educational level).

5.4.1 *Qualitative comparison of questionnaire content*

A comparison of the health domains covered by the 4 health status instruments is presented in Table 5.4.1. The psychological domain is similarly represented in all 4 instruments. The physical domain is operationalized with an emphasis on mobility (NHP, EuroQol) or on overall physical functioning (SF-36, COOP/WONCA). The social (role) domain is underrepresented in the NHP. Social Isolation$_{\text{(NHP)}}$ relates to the ability to make contact with other people, and was thus considered to belong to a psychological rather than a social role domain. The SF-36, the NHP and the EuroQol all contain pain measures. A pain assessment is not included in the standard set of COOP/WONCA charts used in the current study (although a pain chart is optionally available). The EuroQol is the only instrument to address other somatic sensations than pain by combining both 'pain' and 'discomfort' in a single item. The SF-36 is the only instrument to address the concept of positive health (e.g., an item of the Vitality scale referring to feeling 'full of pep'). Despite it's label, the NHP$_{\text{(Energy)}}$ scale contains only negatively worded items (e.g., 'tired all the time', 'everything is an effort'). The NHP is the only measure to assess sleep problems. Both the EuroQol and the SF-36 (the Physical Functioning scale) contain items relating to self-care (e.g., washing, dressing). Both the SF-36 and the COOP/WONCA contain a health transition item (i.e., change in perceived health). All instruments, with the exception of the NHP, provide an assessment of overall health, although this is operationalized in slightly different ways: general health perceptions (SF-36); overall health (COOP/WONCA); or valuation of health (EuroQol).

5.4.2 *Feasibility*

An overview of missing values is presented in Table 5.4.2. The NHP produced the lowest number of missing values. The COOP/WONCA charts, the SF-36 and the EuroQol showed somewhat higher, though acceptable, missing value rates.

Despite the use of appealing pictograms, completing the COOP/WONCA charts was more problematic than expected, with the items Physical Fitness(COOP) (6.3% missing) and Change in Health(COOP) (5.8%) yielding the most missing data. For the EuroQol, the item on Valuation of Own Health had the highest rate of missing data (6.7%).

TABLE 5.4.1 Qualitative comparison of content of NHP, SF-36,COOP/WONCA charts, EuroQol

NHP	SF-36	COOP/WONCA	EUROQOL
Emotional Reactions Social Isolation	Mental Health (MH)	Feelings	Anxiety/Depression
Physical Mobility	--	--	Mobility
--	Physical Functioning (PF)	Physical Fitness	--
--	Role Physical (RP)* + Role Emotional(RE)** + Social functioning (SF)	Daily Activities + Social activities	Usual activities
Pain	Bodily Pain (BP)	--	Pain/Discomfort
--	General Health Perceptions (GH)	Overall Health	Valuation own health
Energy	Vitality (VT)	--	--
--	Reported Health Transition (1 year)	Change in Health (2 weeks)	--
Sleep	--	--	--
--	***	--	Self-care

* Role limitations due to physical health problems
** Role limitations due to emotional problems
*** Physical Functioning(SF-36) contains items relating to self care.

TABLE 5.4.2 Missing values (pooled data)

	range[1]	index[2]
NHP (n=515)	0.4-1.3	0.8
SF-36 (n=1011)	1.1-5.4	3.1
COOP/WONCA (n=496)	0.6-6.3	2.7
EuroQol (n=1011)	3.0-6.7	4.3

[1] range = range in percentage missing values per item
[2] index = (mean number of missing values per respondent / # items) * 100

5.4.3 *Features of score distribution*

Mean scores, standard deviations, and the percentages of respondents with the maximum possible score and the minimum possible score, respectively, for each instrument are shown in Table 5.4.3. The distributions of the scores for all 4 instruments were skewed in the direction of positive health/functioning, as could be expected given the nature of the population under investigation. The EuroQol and the NHP data, with approximately 70-80% of the respondents scoring at the ceiling, were more skewed than the COOP/WONCA and the SF-36 data. The distributions of the EuroQol and the SF-36 compare similarly with those observed in a UK general practice sample (note: the 6D-version of EuroQol).[26]

5.4.4 *Reliability*

The internal consistency coefficients for the SF-36 and NHP scales, based on those respondents who completed both instruments, are shown in Table 5.4.3. The scales of the SF-36 yielded consistently higher internal consistency estimates (mean α= 0.84; range = 0.76 to 0.91) than those of the NHP (mean α= 0.72; range = 0.62 to 0.82). The α-coefficient for 2 of the NHP scales (Energy and Social Isolation) fell below the 0.70 standard recommended for group comparisons. Six of the 8 SF-36 had α-coefficients greater than 0.80.

5.4.5 *Internal structure*

The ICCs for the NHP scales and the SF-36 scales, respectively, are summarized in Table 5.4.5.1 (complete data are shown in Appendix 1). In general, the NHP scales were less highly correlated than were the SF-36 scales. For both instruments, the interscale correlations tended to be higher in the control group than in the migraine group (data not shown). The ICCs between Pain$_{(NHP)}$ and Physical Mobility$_{(NHP)}$ in both the migraine and control groups were remarkably high. Though less pronounced, the same effect was reported for the UK NHP in a general population sample.[27] High ICCs were also observed for Emotional Reaction$_{(NHP)}$ and Social Isolation$_{(NHP)}$, and for Mental Health$_{(SF36)}$ and Social Functioning$_{(SF36)}$.

PCCs between COOP/WONCA items and EuroQol-items, respectively, are summarized in Table 5.4.5.2 (complete data shown in Appendix 2). For both the EuroQol and the COOP/WONCA charts, inter-item correlations were relatively high (summary PCC=0.65 and 0.57 for the EuroQol and COOP/WONCA, respectively). The only exception to this general pattern was the Physical Fitness item of the COOP/WONCA, which exhibited a low correlation with all other items (summary PCC=0.15).

TABLE 5.4.3 Features of score distributions of the NHP (n=515), SF-36 (n=1,011), COOP/WONCA

	mean	S.D.	% max.*	% min.**	Cronbach's α
NHP (score 0-100)					
Physical mobility (8)***	7.2	14.2	70	0	.71
Sleep (5)	11.9	23.1	71	2	.77
Emotional Reactions (9)	10.1	17.3	62	0	.78
Energy (3)	15.6	26.5	70	3	.62
Social Isolation (5)	6.6	15.7	80	0	.63
Pain (8)	7.2	18.7	73	0	.82
SF-36 (score 100-0)					
Physical Functioning (10)	85.5	20.4	38	0	.91
Role physical (4)	70.9	38.7	57	16	.87
Role emotional (3)	78.5	35.8	69	12	.83
Vitality (4)	65.2	18.6	2	0	.79
Mental Health (5)	74.8	18.4	5	0	.87
Social Functioning (2)	81.0	21.1	41	0	.81
Bodily Pain (2)	76.2	22.1	29	0	.88
General Health Perc. (5)	69.6	18.9	4	0	.76
COOP/WONCA (score 1-5)					
Physical Fitness (1)	1.72	1.04	60	3	
Feelings (1)	1.88	.92	40	2	
Daily Activities (1)	1.74	.86	46	1	
Social Activities (1)	1.54	.79	60	1	
Change in Health (1)	2.62	.79	12	1	
Overall Health (1)	2.65	.99	15	2	
EuroQol (score 1-3)					
Mobility (1)	1.15	.38	86	0	
Self care (1)	1.02	.15	98	0	
Usual Activities (1)	1.23	.46	79	2	
Pain/Discomfort (1)	1.43	.54	60	3	
Anxiety/Depression (1)	1.22	.45	80	2	

* % max = percentage of respondents with maximum possible score (ceiling)
** % min = percentage of respondents with minimum possible score (floor)
*** number of items

TABLE 5.4.5.1 Summary* of ICCs for each subscale with the other subscales of NHP resp. SF-36 (pooled data; n=515)

NHP	Energy	Pain	Emot.R	Sleep	Social	Phys.Mob			Overall
	.35	.39	.41	.28	.35	.40			.37
SF-36	PF	RP	RE	VT	MH	SF	BP	GH	Overall
	.38	.41	.37	.46	.47	.52	.48	.47	.45

* For example: the figure of .35 for NHP Energy represents the square root of $(((.33)^2+(.46)^2+(.27)^2+(.35)^2+(.33)^2)/5)$ (Appendix 1)

TABLE 5.4.5.2 Summary* of polychoric correlation coefficients of COOP/WONCA-charts and EuroQol (pooled data; n=496).

COOP/WONCA	Phys	Feel	Daily	Social	Change	Health	Overall
	.15	.58	.78	.64	--	.51	.57
EuroQol	Mob	Self	Usual	Pain	Anxiety		Overall
	.71	.64	.67	.70	.48		.65

* For example: The figure of .58 for Coop/WONCA Feelings represents the square root of $(((.08)^2+(.73)^2+(.72)^2+(.56)^2)/4)$ (Appendix 2)

5.4.6 *Construct validity*

The ICC matrices for NHP scales with SF-36 scales, and the PCC matrices of COOP/WONCA items with EuroQol items are presented in Appendices 1 and 2. The associations observed between the NHP and the SF-36 scales were largely as expected.

Role emotional$_{\text{(SF-36)}}$ correlated best with Emotional Reactions$_{\text{(NHP)}}$; Vitality$_{\text{(SF-36)}}$ with Energy$_{\text{(NHP)}}$; and Mental Health$_{\text{(SF-36)}}$ with Emotional Reactions$_{\text{(NHP)}}$. Energy$_{\text{(NHP)}}$ correlated relatively highly with all SF-36 scales, while the only NHP scale with which Vitality$_{\text{(SF-36)}}$ exhibited a moderate correlation was Energy$_{\text{(NHP)}}$. This latter pattern suggests that Vitality$_{\text{(SF-36)}}$ is a more conceptually distinct scale, while Energy$_{\text{(NHP)}}$ is a more general scale. Physical Functioning$_{\text{(SF-36)}}$ correlated best with Physical Mobility$_{\text{(NHP)}}$ and Pain$_{\text{(NHP)}}$. Role physical$_{\text{(SF-36)}}$ had no counterpart in the NHP.

Feelings$_{\text{(COOP)}}$ correlated best with Anxiety/Depression$_{\text{(EuroQol)}}$, and Usual Activities $_{\text{(EuroQol)}}$ correlated well with Daily Activities$_{\text{(COOP)}}$ and Social Activities$_{\text{(COOP)}}$. Physical Fitness$_{\text{(COOP)}}$ did not correlate well with Mobility$_{\text{(EuroQol)}}$, reflecting the differences in item content. The general nature of Overall Health$_{\text{(COOP)}}$ is evidenced in the fact that it correlated about equally with all of the EuroQol domains.

Common factor analysis of the combined data of the SF-36 (scales), COOP/WONCA (items) and EuroQol (items) yielded two factors with an eigenvalue >1.0, together explaining 52% of the common variance. The factor loadings after varimax rotation are shown on the left-side of Table 5.4.6. The first factor extracted appears to reflect a mental health dimension; the second factor a physical health dimension.

Common factor analysis of the combined data of the NHP (scales), SF-36 (scales) and EuroQol (items) yielded two similar factors, together explaining 54% of the common variance (see Table 5.4.6, right-hand side).

5.4.7 Known Groups Validity

Table 5.4.7 reports the data relating to the ability of the 4 health status measures to discriminate between 'known groups' characterized by differences in: (1) the number of self-reported chronic conditions in the previous year; and (2) the number of days absent from work due to illness in the 2 weeks prior to assessment. The SF-36 scales discriminated best between the groups reporting one or less versus more than one chronic condition. All p-values were beyond <.01, with effect sizes being in the moderate to high range (.49 to .92). The NHP scales also discriminated clearly between these groups, with the effect sizes being in the moderate range (d around .50).

Five of the 6 COOP/WONCA charts yielded significant group differences (with the exception of Physical Fitness), with effect sizes ranging from .12 to .74. Statistically significant group differences were also observed for all of the EuroQol items, although the effect sizes were more variable (ranging from .13 for Self-Care to .84 for Pain/Discomfort).

The parallel analyses employing absence from work as the grouping variable yielded similar results. The SF-36 performed best, with all scales yielding statistically significant group differences, and 7 of the 8 scales exhibiting large effect sizes (d > .80). The very high effect size estimate for Role physical(SF-36) is probably at least in part an artifact of the conceptual overlap between the criterion 'absence from work' and the content of Role physical (SF-36). This makes the high effect size estimates of the other SF-36 scales, which do not have a high degree of conceptual overlap with the criterion (Bodily Pain(SF-36), Social Functioning(SF-36), Vitality(SF-36), Physical Functioning(SF-36)) all the more striking. The NHP yielded consistently significant results, although the effect sizes tended to be lower than in the analysis of chronic disease groups (d ranging from .47 to .74). Only 4 of the 6 COOP/WONCA charts yielded statistically significant group differences, with effect sizes ranging from .71 to 1.02. All of the EuroQol items discriminated clearly between the two groups, with effect sizes ranging from .54 to 1.16.

TABLE 5.4.6 Common factor analyses of the SF-36 (scales), EuroQol (items), and COOP/WONCA (items) [left; n=496]; and of SF-36 (scales), EuroQol (items) and NHP (scales) [right; n=515]; factor loadings > 0.3 after varimax rotation.

	Factor 1	Factor 2	Factor 1	Factor 2
% of common variance explained	*29*	*23*	*28*	*26*
COOP/WONCA				
Physical Fitness	--	--		
Feelings	.85	--		
Daily Activities	.69	.51		
Social Activities	.71	--		
Change in Health	--	--		
Overall Health	.54	.53		
SF-36				
Physical Functioning	--	.83	.87	--
Role physical	.41	.59	.53	.42
Role emotional	.69	--	--	.70
Vitality	.70	--	--	.71
Mental Health	.88	--	--	.86
Social Functioning	.69	.42	.42	.64
Bodily pain	--	.66	.66	--
General health perceptions	.46	.56	.53	.49
EuroQol				
Mobility	--	.64	.66	--
Self care	--	.41	.46	--
Usual Activities	--	.68	.66	--
Pain/Discomfort	--	.62	.62	--
Anxiety/Depression	.70	--	--	.71
NHP				
Energy			.45	.51
Pain			.81	--
Emotional Reactions			--	.83
Sleep			--	--
Social Isolation			--	.65
Physical Mobility			.86	--

TABLE 5.4.7 Discriminative ability of NHP, SF-36, COOP/WONCA and EuroQol between groups differing in the number of self-reported chronic conditions; and the number of days of absence due to illness in the two weeks prior to assessment, respectively.

	Self-reported chronic conditions				Absence from work due to illness			
	< 1 n=408[1] X (s.d.)	> 1 n=603[1] X (s.d.)	p-value (MWU)	Effect size (*d*)[2]	0 days n=395[1] X (s.d.)	> .5 days n=66[1] X (s.d.)	p-value (MWU)	Effect size (*d*)[2]
NHP (range 0-100)								
Energy	5.1 (14.9)	22.4 (29.9)	< 0.01	.69*	11.2 (21.2)	22.2 (30.6)	< 0.01	.48
Pain	1.9 (7.7)	13.2 (22.0)	< 0.01	.63*	4.6 (11.1)	11.4 (19.5)	.03	.54*
Emotional Reactions	3.5 (8.8)	14.3 (20.0)	< 0.01	.65*	8.4 (13.9)	15.2 (17.2)	< 0.01	.47
Sleep	5.3 (14.3)	16.1 (26.4)	< 0.01	.47	6.3 (15.0)	17.6 (26.6)	< 0.01	.66*
Social Isolation	2.4 (8.2)	9.2 (18.5)	< 0.01	.44	3.5 (11.6)	9.1 (13.2)	< 0.01	.47
Physical mobility	2.3 (7.8)	10.4 (16.4)	< 0.01	.59*	3.3 (7.5)	10.2 (16.5)	< 0.01	.74*
SF-36 (range 100-0)								
Physical functioning	92.8 (13.2)	80.4 (22.8)	< 0.01	.64*	92.5 (12.1)	78.1 (24.2)	< 0.01	.99**
Role physical	85.4 (29.7)	60.8 (41.0)	< 0.01	.67*	80.2 (32.5)	36.0 (38.5)	< 0.01	1.35**
Role emotional	88.7 (26.8)	71.6 (39.4)	< 0.01	.49	82.5 (32.2)	66.7 (40.5)	< 0.01	.47
Vitality	72.1 (15.7)	60.5 (19.0)	< 0.01	.65*	68.9 (15.4)	53.2 (21.4)	< 0.01	.96**
Mental health	81.3 (14.1)	70.4 (19.7)	< 0.01	.62*	78.2 (14.9)	65.7 (19.6)	< 0.01	.80**
Social functioning	89.2 (17.1)	75.4 (21.8)	< 0.01	.64*	85.7 (16.4)	64.6 (23.8)	< 0.01	1.19**
Bodily pain	87.2 (16.9)	68.6 (22.1)	< 0.01	.92**	81.2 (18.1)	58 2 (24.6)	< 0.01	1.20**
General health perceptions	77.1 (14.7)	64.4 (19.8)	< 0.01	.71*	75.0 (15.3)	59.2 (21.4)	< 0.01	.91**
COOP/WONCA (range 1-5)								
Physical Fitness	1.68 (.97)	1.75 (1.09)	.73	.06	1.59 (.91)	1.91 (1.29)	.26	.33
Feelings	1.63 (.68)	2.05 (1.02)	< 0.01	.47	1.77 (.77)	2.45 (1.21)	< 0.01	.80**
Daily Activities	1.46 (.69)	1.94 (.89)	< 0.01	.59*	1.56 (.63)	2.24 (1.11)	< 0.01	.94**
Social Activities	1.33 (.63)	1.70 (.86)	< 0.01	.49	1.40 (.61)	1.85 (.75)	< 0.01	.71*
Change in health	2.68 (.70)	2.57 (.83)	< 0.01	.12	2.57 (.76)	2.73 (1.03)	.15	.20
Overall health	2.25 (.95)	2.94 (.91)	< 0.01	.74*	2.36 (.87)	3.27 (1.00)	< 0.01	1.02**
EuroQol (range 1-3)								
Mobility	1.05 (.23)	1.22 (.43)	< 0.01	.50*	1.06 (.24)	1.24 (.53)	< 0.01	.60*
Self-care	1.01 (.10)	1.03 (.18)	.03	.13	1.00 (0.00)	1.08 (.28)	< 0.01	.80**
Usual activities	1.10 (.36)	1.32 (.50)	< 0.01	.50*	1.11 (.32)	1.55 (.61)	< 0.01	1.16**
Pain/discomfort	1.18 (.40)	1.60 (.56)	< 0.01	.84**	1.31 (.47)	1.62 (.58)	< 0.01	.63*
Anxiety/depression	1.08 (.30)	1.31 (.51)	< 0.01	.50*	1.15 (.36)	1.36 (.54)	< 0.01	.54*

[1] Due to the study design we had different numbers of cases available for NHP, COOP/WONCA and SF-36 + EuroQol, respectively. The numbers of NHP- and COOP/WONCA cases were weighted up to the sample size of SF-36 and EuroQol.

[2] Interpretation: *d* = .2: small effect; *d* = .5: medium effect; *d* = .8: large effect.

* .5 < *d* < .8

** *d* > .8

5.5 Discussion

In this study we have compared the feasibility, structure and psychometric characteristics of 4 well-known generic health status measures - the NHP, the SF-36, the COOP/ WONCA Charts, and the EuroQol - when employed in a large sample of migraine sufferers and a matched control group from the general population.

Despite inherent differences in their design (e.g., multi-item scales versus single-item measures; dichotomous versus categorical response choices), broadly speaking, all 4 of these measures address two basic health domains: physical and mental health and functioning. The qualitative comparison of these measures, however, indicates that each approaches the topic areas covered from a somewhat different perspective. For example, despite the similarity in labels, the NHP 'Energy' scale and the SF-36 'Vitality' scale differ in the type and range of subjective health experiences elicited from respondents. While the NHP focuses on symptoms of fatigue, the SF-36 includes a mix of both positive and negative items. Similarly, while the EuroQol mental health item focuses on anxiety and depression, the SF-36 'Mental Health' scale includes positive emotions as well (e.g., feeling 'calm and peaceful').

The feasibility of the measures (i.e., the ease with which they can be completed by respondents) was examined indirectly by calculating rates of missing values. Importantly, the length of an instrument does not appear to have any direct bearing on the frequency of missing responses. For example, the highest rate of missing values was observed for the EuroQol, one of the shortest of the instruments investigated. The NHP had the lowest missing value rate; lower than the proportions of missing data reported for the UK version of the instrument.[27] This reflects the simple, dichotomous response choices used consistently throughout the questionnaire, as well as the low demands placed on the respondents' reading skills via the use of short, uncomplicated sentences. Although not examined in the current study, the simplicity of the item wording and response choices of the NHP, combined with its negative question valence, may also make it susceptible to acquiescence response sets.[19]

Interestingly, the use of visual aids (i.e., pictograms) in the COOP/WONCA charts does not necessarily guard against respondent errors. In fact, one of the charts (Physical Fitness) yielded the second highest missing value rate across measures. The missing value rates for the SF-36 observed in the current study were comparable to those reported for the UK and the US versions of the instrument.[16,27] It should be noted that, because the sequencing of the 4 questionnaires was varied, an ordering effect cannot account for the observed differences between the instruments in rates of missing values.

The SF-36 scales exhibited high levels of internal consistency, which are comparable to those reported for the US version of the questionnaire when employed in a sample of patients with chronic health conditions (αs ranging from 0.78 to 0.93)[16], and in a sample of migraine sufferers (αs ranging between 0.80 and .88).[28] Lower, though generally acceptable reliability estimates were found for the NHP scales. The differences in scale reliabilities noted between the SF-36 and the NHP may be due, in part, to the type of data generated by the two instruments (i.e., the SF-36 yields polytomous data, the NHP dichotomous data). It might be argued that the NHP sacrifices some internal consistency for the sake of simplicity.

Examination of the inter-scale correlations (ICCs) for the NHP and for the SF-36 indicated that the NHP scales are somewhat less highly correlated with one another than are those of the SF-36. In general, however, the inter-scale correlations for both the NHP and the SF-36 were of a low to moderate magnitude, suggesting little redundancy in the type of information generated by the various scales with these two instruments.

In contrast, the COOP/WONCA charts were found to be relatively highly correlated with one another (overall PCC=0.57) as were the items of the EuroQol (overall PCC=0.65). This suggests that there is a higher degree of conceptual overlap among the items in these two instruments than is the case with the multi-item scales of the NHP and SF-36, respectively.

The pattern of ICCs observed *between* the scales of the NHP and the SF-36 was generally consistent with expectations, with conceptually similar scales yielding the highest correlations. One exception was the relatively low correlation observed between the Pain$_{(SF-36)}$ and Pain$_{(NHP)}$ scales in the migraine group (ICC = .46; comparable ICC in the control group .65). This may be explained by the fact that the SF-36 pain items refer to bodily pain, in general, whereas the corresponding NHP items focus on pain as it relates specifically to physical movement. Thus, while migraine sufferers might report more pain on the Bodily Pain$_{(SF-36)}$ scale, this would not necessarily be the case for the Pain$_{(NHP)}$ scale.

The overall pattern of correlations *between* the COOP/WONCA charts and the EuroQol items was also consistent with expectations. Particularly striking was the generally high correlation between the Overall Health$_{(COOP)}$ and all of the EuroQol items (all PCCs >0.60), suggesting that this COOP/WONCA chart is indeed tapping a general health construct.

The two combined factor analyses (i.e., the SF-36 scales, the COOP/WONCA and the EuroQol items; the SF-36 and NHP scales, and the EuroQol items) yielded remarkably similar results. Both analyses resulted in an intuitively appealing solution, with two higher-order factors being identified - one reflecting physical health; the other mental health. A similar 2-dimensional model has been proposed for the Sickness Impact Profile (SIP)[3] and the Rosser-Kind Matrix.[29] In the SIP, scores for 3 (out of 12) scales are combined to describe physical dysfunction, scores for 4 other scales are combined to describe psychosocial dysfunction, while the remaining 5 scales are named ‘independent’. The Rosser-Kind Matrix consists of a classification of illness along two dimensions ‘disability’ and ‘distress’, respectively. The results of our study are similar to those obtained in other factor analytic studies of the MOS measures and of the SIP.[10,30] If these results are replicated in future studies, it may be possible to efficiently summarize health status data by physical and mental health component scores. This could increase the precision of such scores, and could facilitate certain types of studies (e.g., cost-effectiveness) which require the use of summary health status indicators.

The tests of ‘known groups’ validity indicated that, of the 4 instruments examined, the SF-36 was best able to discriminate between groups formed on the basis of chronic disease status, and work disability days. Of the 3 remaining instruments, the NHP and the EuroQol also performed well, while 4 out of the 6 COOP/WONCA charts evidenced good discriminative ability. These results confirm earlier reports of the ability of the SF-36 to discriminate between patients with minor versus major medical conditions, and between patients with physical health versus psychiatric conditions.[9] It should be noted that evidence of discriminative power based on cross-sectional analyses does not necessarily imply that an instrument will also be responsive to changes in health status over time. Although this may well be the case, this needs to be confirmed empirically with the use of longitudinal study designs.

An overall summary of the results of this study is reported in Table 5.5. All 4 of the health status instruments examined yielded low levels of missing data. This finding adds to the already substantial body of evidence supporting the feasibility of collecting subjective health status data in relatively large scale survey research settings. The question of whether similar assessments can be successfully incorporated into more clinically-oriented, longitudinal studies of seriously ill patient populations is the subject of current study.

In general, all 4 instruments exhibited a good performance profile, including reliability (where assessed), construct validity, and 'known groups' validity. Of the two health profiles investigated, the SF-36 performed best psychometrically, exhibiting highest scale internal consistency and discriminative ability. Both health profiles outperformed the COOP/WONCA charts and the EuroQol, reflecting the psychometric advantages often associated with instruments having a multi-item scale structure.

TABLE 5.5 Summary of empirical comparison of NHP, SF-36, COOP/WONCA charts and EuroQol

	NHP	SF-36	COOP/WONCA	EuroQol
Missing value rate	best	acceptable	acceptable	acceptable
Internal consistency	acceptable	best	*not applicable*	*not applicable*
Construct validity	confirmed	confirmed	confirmed	confirmed
Discriminative ability	good	best	good	good

Additional research is needed to provide a head-to-head comparison of the test-retest reliability of these instruments, as well as of other aspects of their validity, including particularly their responsiveness to change in health status over time. Additionally, more formal, confirmatory tests are needed (e.g., using structural equation models) to explore further the underlying, higher-order physical and mental health score components identified in the current study. Finally, the relative performance of these measures when employed with more seriously ill patient populations needs to be further investigated.

Ultimately, choosing among available generic health status instruments requires not only a careful consideration of their formal psychometric properties, but also of the match between their substantive content (e.g., the breadth and depth with which they address relevant health domains) and the specific research question at hand. Additionally, practical considerations such as respondent burden, and the availability of culturally- and language-adapted versions can be important in identifying the most appropriate measure for use in a given study. Finally, it may be imprudent to approach such decisions from an 'either-or' perspective. The use of several generic measures, or combining generic with disease-specific measures in a single study, may yield the greatest return on our investment in health status assessment. Particularly given that many of the available generic instruments are quite brief, such a strategy should be possible without resulting in excessive respondent burden.

Notes

1. Separate factor analyses for each of the 4 health status measures were also carried out. The results were similar to those derived from the two analyses in which 3 of the 4 instruments were examined simultaneously. Results of these additional analyses are available upon request.

2. Given unequal score variance between groups, the denominator used in calculating the d statistic was the square root of:
   $[(N_a - 1)S_a^2 + (N_b - 1)S_b^2] / [(N_a - 1) + (N_b - 1)]$

Acknowledgement

The authors wish to thank the respondents who participated in this study, Arjan P Bandel and Leona van Royen, MSc for their contributions to the research underlying this paper, and Paul J van der Maas, MD, PhD and John E Ware Jr, PhD for their valuable comments on earlier drafts of this paper. We are also grateful for information provided by our colleagues associated with the Dutch Working Group on Health Status Assessment.

APPENDIX 1 Intraclass correlation coefficients of NHP-scales and SF-36 scales* (pooled data, n=515).

	Energy	Pain	Emotion	Sleep	Social	Phys Mob	PF	RP	RE	VT	MH	SF	BP
Pain	.33												
Emotion	.46	.23											
Sleep	.27	.32	.34										
Social	.35	.16	.62	.22									
Phys.Mob	.33	.71	.26	.27	.23								
PF	.44	.69	.28	.32	.19	.67							
RP	.36	.23	.21	.23	.16	20	.33						
RE	.35	.13	.46	.16	.29	.14	.20	.42					
VT	.47	.17	.29	.20	.20	.16	.29	.40	.33				
MH	.41	.18	.56	.24	.35	.18	.27	.28	.52	.59			
SF	.52	.32	.48	.29	.32	.33	.46	.50	.47	.50	.63		
BP	.46	.43	.29	.29	.16	.37	.58	.53	.23	.42	.36	.60	
GH	.41	.28	.27	.26	.17	.23	.41	.39	.31	.60	.49	.49	.53

* As NHP and SF-36 scales run in the opposite direction, we used (100 - (SF-36 scale score)) in the determination of ICCs.
ENERGY = NHP Energy; PAIN = NHP Pain; EMOTION = NHP Emotional Reaction; SLEEP = NHP Sleep; SOCIAL = NHP Social Isolation; PHYSMOB = NHP Physical Mobility
PF = SF-36 Physical functioning; RP = SF-36 Role physical; RE = SF-36 Role emotional; VT = SF-36 Vitality; MH = SF-36 Mental health; SF = SF-36 Social functioning; BP = SF-36 Bodily pain; GH = SF-36 General health perceptions.

APPENDIX 2 Polychoric correlations of COOP/WONCA-items and EuroQol-items (pooled data, n=496).

	EQ-Mob	EQ-Self	EQ-Usual	EQ-Pain	EQ-Mood	CO-Phys	CO-Feel	CO-Daily	CO-Soc	CO-Change
EQ-Self	.70									
EQ-Usual	.71	–*								
EQ-Pain	–*	.77	–*							
EQ-Mood	–*	.40	.51	.52						
CO-Phys	.26	.56	.33	.19	–*					
CO-Feel	.19	.38	.46	.39	.83	.08				
CO-Daily	.58	.79	.75	.64	.71	–*	.73			
CO-Social	.42	.62	.61	.52	.68	.10	.72	0.82		
CO-Change	–*	.10	–*	–*	–*	–*	–*	–*	–*	
CO-Overall	.61	.69	.79	.66	.68	.22	.56	–*	.66	–*

–* polychoric correlation coefficient is unreliable, because the assumption of a bivariate normal distribution of data is not fulfilled.
EQ-MOB = EuroQol Mobility; EQ-SELF = EuroQol Self-care; EQ-USUAL = EuroQol Usual Activities; EQ-Pain = EuroQol Pain/discomfort; EQ-MOOD = EuroQol Anxiety/depression.
CO-PHYS = COOP/WONCA Physical Fitness; CO-FEEL = COOP/WONCA Feelings; CO-DAILY = COOP/WONCA Daily activities; CO-SOCIAL = COOP/WONCA Social Activities; CO-CHANGE = COOP/WONCA Change in health; CO-OVERALL = COOP/ WONCA Overall health

References

1 Ware JE. *Generic versus specific health status measures.* Medical Outcomes Trust Bulletin 1994;2:4.

2 World Bank. *Investing in health: world development indicators.* World Development Report 1993. Oxford: Oxford University Press, 1993.

3 Bergner M, Bobbitt RA, Carter WB, Gilson BS. *The Sickness Impact Profile: development and final revision of a health status measure.* Med Care 1981;19:787-805.

4 Hunt S, McEwen J, McKenna SP. *Measuring Health Status.* London: Croom Helm, 1986.

5 Weel C. van. *Functional status in primary care: COOP/WONCA charts.* Disability and Rehabilitation 1993;15:96-101.

6 EuroQol Group. *EuroQol - a new facility for the measurement of health-related quality of life.* Health Policy 1990;16:199-206.

7 Essink-Bot ML, Stouthard MEA, Bonsel GJ. *Generalizability of valuations on health states collected with the Euroqol questionnaire.* Health Economics 1993;2:237-246.

8 Ware JE, Sherbourne CD. *The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection.* Med Care 1992;30:473-481.

9 McHorney CA, Ware JE, Raczek AE. *The MOS 36-item Short Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs.* Med Care 1993;31:247-263.

10 Stewart AL, Ware JE. *Measuring functioning and well-being: the Medical Outcomes Study approach.* Durham/London, Duke University Press, 1992.

11 Royen L van, Essink-Bot ML, Koopmanschap MA, Michel BC, Rutten FFH. *A society's perspective on the burden of migraine in the Netherlands.* PharmacoEconomics 1995;7 (2):170-179.

12 Essink-Bot ML, Royen L van, Krabbe PFM, Bonsel GJ, Rutten FFH. *The impact of migraine on health status.* Headache 1995;35(4):200-206.

13 Headache Classification Committee of the International Headache Society. *Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain.* Cephalalgia 1988;8(Suppl 7):1-96.

14 Bonsel GJ, Essink-Bot ML, Klompmaker IJ, Slooff MJH. *Assessment of the quality of life before and following liver transplantation.* Transplantation 1992;53:796-800.

15 Essink-Bot ML, Krabbe PFM, Agt HME van, Bonsel GJ. *NHP or SIP: a comparative study in renal insufficiency associated anemia.* In press in Quality of Life Research.

16 McHorney CA, Ware JE, Lu JFR, Sherbourne CD. *The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse*

patient groups. Med Care 1994;32:40-66.

17 Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, Bungay K, et al. *International quality of life assessment (IQOLA) project.* Quality of Life Research 1992;1:349-351.

18 Scholten JHG, Weel C van. *Functional status assessment in family practice. The Dartmouth COOP Functional Health Assessment Charts/WONCA.* Lelystad, The Netherlands: Meditekst, 1992.

19 Streiner DL, Norman GR. *Health measurement scales - a practical guide to their development and use.* Oxford: Oxford Medical Publications, 1989.

20 Nunnally JC. *Psychometric theory.* New York: McGraw Hill, 1978.

21 Dunn G. *Design and analysis of reliability studies.* Oxford: Oxford University Press, 1989.

22 Deyo RA, Diehr P, Patrick DL. *Reproducibility and responsiveness of health status measures.* Controlled Clinical Trials 1991;12:142S-158S.

23 Crocker L, Algina J. *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston, Inc. 1986.

24 Jöreskog KG, Sörbom D. PRELIS. *A preprocessor for LISREL.* Mooresville, USA: Scientific Software, Inc. 1986.

25 Cohen J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1977.

26 Brazier J, Jones N, Kind P. *Testing the validity of the EuroQol and comparing it with the SF-36 health survey questionnaire.* Quality of Life Research 1993;2:169-180.

27 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. *Validating the SF-36 health survey questionnaire: new outcome measure for primary care.* BMJ 1992;305:160-164.

28 Osterhaus JT, Townsend RJ, Gandek B, Ware JE. *Measuring the functional status and well-being of patients with migraine.* Headache 1994;337-343.

29 Rosser R, Kind P. *A scale of valuations of states of illness: is there a social consensus?* Int J Epidemiol 1978;7:347-358.

30 Bruin AF de, Diederiks JPM, Witte LP de, Stevens FCJ, Philipsen H. *The development of a short generic version of the Sickness Impact Profile.* J Clin Epidemiol 1994;47:407-418.

6

The impact of migraine on health status

Published as:
Essink-Bot ML, Roijen L van, Krabbe PFM, Bonsel GJ, Rutten FFH. The impact of migraine on health status. Headache 1995;35(4):200-206. Reprinted with permission of the publisher.

6.1 Abstract

Problems. What is the effect of migraine on health status, defined as the subject's physical, psychological and social functioning? And, suppose that the health status of migraine sufferers appears to be impaired, to what extent is this a consequence of migraine-associated comorbidity rather than of migraine itself?
Methods. A group of 846 migraineurs, selected from the general population following IHS criteria, and a control group were surveyed with the Short Form-36, the Nottingham Health Profile, the EuroQol instrument and the COOP/WONCA charts. Questions on demographic characteristics and comorbidity were included.
Results. The health status of migraineurs appeared to be significantly impaired in comparison to the control group. Because statistical significance is distinct from relevance, effect size estimators were employed. Although the direction of the differences indicated consistently a worse health status of the migraineurs, regardless of the instrument used, the sizes of the differences were small to medium. Self-reported comorbidity, especially depression, was more prevalent in the migraine group. However, this offered only a partial explanation for the impaired health status of the migraine group.
Conclusions. Migraine has an independent, moderately deteriorating effect on the daily functioning of individuals.

6.2 Introduction

The burden of migraine, a chronic, attack-wise and presumably disabling disease, should not be underestimated. The reported one-year prevalence in adults exceeds 10%, with a male to female ratio of about 1 to 2-3.[1] People in the age range 15-55 years are predominantly afflicted, i.e., those in the work force. Long term consequences of migraine may result from interference of frequent attacks with daily life, thus precluding optimal functioning. We designed a study to quantify the burden of migraine both in terms of its economic consequences and in terms of its impact on health status.

The economic part of the study, that was published in detail elsewhere, showed that direct costs of migraine in the Netherlands accounted for 0.3% of the total health care costs in 1988, 80% of which could be attributed to 'alternative' medical practice. Indirect costs, due to absence from work and reduced productivity, were estimated to amount to at least 542 million Dutch guilders per year (1988 1$ = 1.9 *f*).[2]

Health status, the focus of the present paper, is defined as physical, psychological and social functioning. Osterhaus concluded from a survey of 845 migraineurs (meeting IHS criteria[3]) with the Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) that 'although migraineurs may be physically able to function, they function behaviorally at a level well below their physical capabilities, and for some domains even worse than patients suffering from arthritis, gastrointestinal disorders or diabetes'.[4] Solomon assessed the health status of 208 patients attending a Headache Center with the Medical Outcomes Study 20-item instrument (MOS-20). The authors conclude 'that chronic headache disorders are associated with significant limitations in all measured dimensions of patient well-being and functioning when compared to patients with no chronic condition; and that patients with chronic headaches have a level of function worse than that of patients suffering from diabetes, arthritis, depression and back problems'.[5] Jenkinson reported the results of interviewing 80 women attending a migraine out-patient clinic (diagnosed as suffering from migraine by a neurologist) with the Nottingham Health Profile and the General Health Questionnaire (GHQ; a screening instrument for non-psychotic psychiatric disturbance).[6] GHQ-scores were indicative of mood disturbance in no less than 41% of the subjects. In a Dutch study among elderly patients (age range 55 - 79 years) only 9% of those who stated they suffered from 'migraine or severe headache' reported physical limitations, and 10% gave a negative evaluation of their general health. However, 45% regarded their psychological well-being as being impaired by their headache complaints.[7] Overall, these studies are indicative of a worse functioning of migraine sufferers. However, controlled studies, enabling a comparison between migraine sufferers and non-afflicted subjects and an estimation of the size of the effect of migraine on health status, are not known to us.

It has been recognized that migraine often occurs in association with other conditions, like mood disturbances (depression, anxiety)[8,9,10], allergic phenomena (atopy, asthma, food allergy)[11,12] and vasospastic disorders (Raynaud's phenomenon).[13,14,15] This higher prevalence of comorbidity was confirmed in a recent Dutch survey on socio-economic health inequalities in a representative sample of the general population (n=15,973; age range 15-64). The prevalence of self-reported migraine (no check on IHS criteria) was 12% for women and 5% for men. Women with migraine reported no other chronic condition in 39% of cases, while 15% reported 2 or more; for women without migraine these figures are 60% and 5%. The largest difference in prevalence of a specified chronic condition was for 'depression/nervous exhaustion' (22% for women with migraine, 6% for women without migraine). Similar figures held for men with and without migraine in this study. (K Stronks, Departement of Public Health, Erasmus University Rotterdam; personal communication, 1994).
With regard to the *causal* relationship between migraine and comorbid conditions several authors have proposed a common disposition or a common pathogenetic

defect.[12,16] Information about the *consequences* of the higher prevalence of comorbidity in migraine sufferers is scarce. In particular, the relative contribution of migraine and other conditions to the lower level of functioning by migraine patients has not been investigated previously.

In the present study the health status of migraine patients is compared with that of a control group. We intend to answer the following questions:

1. What is the health status of migraine sufferers compared with a control group that is comparable on age, gender and employment status?
2. Are the differences between migraineurs and controls consistent if measured with different generic instruments?
3. What is the relative contribution of migraine and associated comorbidity, especially self-reported depressive disorders, to the impaired health status of migraine sufferers?

6.3 Methods

6.3.1 Samples

Migraine patients were selected from a series of face-to-face interviews with a representative sample of the Dutch general population (n=10,480), avoiding the selection of only severe cases who seek medical care, during the period October 1992 to February 1993. Subjects were included as migraine patients if they met the IHS criteria[2] *and* had experienced at least one attack of migraine during the 12 months prior to the interview. 992 migraine sufferers met these criteria (1-year prevalence 9.5%). Of these sufferers, who were all invited to participate in a second study, i.e., the actual investigation on health status and (in)direct costs, 85% (n=846) actually agreed to cooperate.

The control group was selected from the subjects in the survey who did not meet the criteria for migraine by frequency matching to the migraine group on 5-year age class, gender and employment status.

6.3.2 Instruments

Generic instruments for health status assessment measure basic values (physical, psychological and social functioning) which are relevant for everyone's health status.[17] There is general agreement that the primary source for such information is to be found in the subjects themselves. Generic questionnaires are non-disease specific, enabling comparison of health status data across the borders of specified diagnoses.

A combination of four generic questionnaires, the MOS Short Form-36, the Nottingham Health Profile (Dutch Adaptation), the EuroQol descriptive instrument and the COOP/WONCA charts, was applied to investigate whether differences between migraineurs and controls were consistent if measured with different instruments. Data were also analysed to compare testing properties of these questionnaires.

The *Medical Outcomes Study 36-item Short-Form Health Survey (SF-36)* was developed in the US from the Medical Outcome Study General Health Survey Instrument.[18,19,20] It consists of 36 items, assigned to the domains of Physical Functioning (10 items), Social functioning (2), Role limitations (physical problems) (4), Role limitations (emotional problems) (3), Mental Health (5), Vitality (4), Pain (2), General

Health Perceptions (5) and Health Change (1). The numbers of response categories per item range from 2 to 6. The end score is an eight-dimensional profile. The Dutch version we used was developed as a part of the IQOLA project, which aims to translate, validate and norm the SF-36 in a range of languages and cultural settings.[21]

The *Nottingham Health Profile (NHP)* was developed during the seventies in the UK as a measure for perceived health, to be used in population surveys.[22] Part 1 of the NHP consists of 38 dichotomous items, covering the domains of Physical Mobility (8 items), Pain (8), Energy (3), Sleep (5), Social Isolation (5) and Emotional Reaction (9). Part 2 consists of seven items on problems because of health in seven specified areas of life. The Dutch version we used has been tested in several patient populations. [23,24]

The *EuroQol* classification consists of five items (Mobility, Self-care, Usual activities, Pain/discomfort and Anxiety/depression), each following the general form: no problems - some problems - extreme problems.[25] Additionally, evaluation of own health is assessed with a visual analogue scale ranging from 0 (worst imaginable health state) to 100 (best imaginable health state). The EuroQol instrument was developed by the international EuroQol Group as a standardised, non-disease-specific measure for description of health status. EuroQol health state descriptions can be linked directly to empirical valuations of health states by the general population, a feature which makes it especially interesting for the economic assessment of medical interventions.

The *COOP/WONCA charts* were developed to assess health status of patients in primary care.[26] There are six charts, covering the domains of Physical Fitness, Feelings, Daily Activities, Social Activities, Change in Health and Overall Health. The levels on the scales are illustrated with pictograms.

Comorbidity was assessed by the list of chronic conditions, as included in the Dutch Health Interview Survey of the Dutch Central Bureau of Statistics. This list counts 28 conditions in lay terms (like 'asthma, chronic bronchitis or COPD', 'diabetes', 'varicose veins'). Respondents are asked to indicate for each condition whether they have it now or if they have had it in the year prior to assessment.

6.3.3 Questionnaire lay-out and mailing scheme

We used four different questionnaires, two for the migraine group and two for the control group. All versions contained the SF-36, EuroQol and questions relating to comorbidity and demography. The two migraine versions differed from each other, one containing the COOP/WONCA charts, the other the NHP. The two control group versions differed in the same way. Both migraine versions contained additional questions on the number of attacks during the year prior to assessment and on medical consumption.

Questionnaires were sent by mail in June, 1993, with reminders two weeks (a postcard) and five weeks (a complete questionnaire) later.

6.3.4 Analysis

To investigate any selectivity of response, non-response analyses were conducted by comparing and testing (Chi-square test) the distributions of age, gender, social class and degree of urbanization of addressees and respondents.

Scores were declared as missing values if nothing was filled in or if ambiguous information was provided. Because of generally low missing value rates we did not impute constructed values for missings. Scale scores for the SF-36 and NHP were based on complete records only.

The Mann-Whitney U test was applied for testing differences in scores of continuous non-normally distributed variables between the migraine group and controls. To avoid the effect of multiple testing, $p < 0.01$ was regarded as statistically significant. Given the large sample size, statistical significance may be misleading: relatively small mean differences will achieve conventional levels of statistical significance without representing meaningful differences in functioning. We employed an estimator of effect size d for continuous variables, which relates the differences in mean scores to the dispersion of the scores. A $d = .2$ indicates a small effect, a $d = .5$ a medium effect and a $d = .8$ a large effect.[27]

The Chi-square test was used to test for proportional differences in contingency tables. Again, $p < 0.01$ was regarded as statistically significant. The effect size estimator W for contingency tables has a different interpretation: $W = .1$ indicates a small effect, $W = .3$ a medium effect, $W = .5$ a large effect.[27]

Multiple classification analysis (MCA) was applied to explore the relative effects on health status of migraine and associated comorbidity.[28,29] Essentially, MCA is multiple regression analysis using dichotomous predictor (or explanatory) variables. We used 'migraine yes/no', 'depression yes/no' and 'diseases of the skin yes/no' as predictor variables. The choice of the latter two conditions was based on significant differences of their prevalences in the migraine group and the control group. The scale scores of the SF-36, NHP and EuroQol (valuation of own health) that showed the largest differences between the migraine group and the control group were used as dependent (or explained) variables in separate MCAs.

It can be argued that loglinear analysis would be more appropriate, because for MCA a continous and normal distribution of the dependent variable is required. Application of loglinear analysis did not change the conclusions. We have chosen to present MCA results as they are easier to interpret.

6.4 Results

6.4.1 Response

The questionnaire was mailed to 846 migraine sufferers as identified by the diagnostic interview. 65 of them returned it, remarking they did not have migraine. A number of migraineurs as classified by the diagnostic interview probably did not label their headaches as migraine themselves. After exclusion of these 65 and after correction for wrong addresses, the crude response-rate was 63%. Of these, 90% were usable (n=436). There were no significant differences in response rates between the two migraine groups (questionnaire with COOP/WONCA charts or NHP respectively). 843 questionnaires were mailed to the control group. After correction for wrong addresses, the crude response rate was 72%. All but ten were usable (n=575). As in the migraine group, there were no significant differences in response rates between the two control groups.

Due to the different composition of the questionnaires, the following numbers per instrument were available for analysis: for SF-36 and EuroQol, n=436 in the migraine group and n=575 in the control group; for NHP, n=226 and n=289; for COOP/ WONCA n=210 and n=286.

The non-response analyses did not show significant differences between addressees and respondents in either the migraine group or the control group, suggesting no selective non-response.

6.4.2 Respondents' characteristics

Demographic characteristics and data relating to the prevalence of self-reported comorbidity are presented in Table 6.4.2. The differences between the respondents in the migraine group and the controls were not significant for sex distribution, age, employment status or educational level. However, after exclusion of 'migraine' and 'severe headache', the respondents in the migraine group reported significantly more chronic conditions now or in the past year. Especially 'diseases of the skin/eczema' and 'depression/ nervous exhaustion' were more prevalent in the migraine population (14% and 29% in the migraine group, 9% and 16% in the control group respectively).

The migraine patients reported an average number of 13 attacks of migraine during the past twelve months (41%, 4 or fewer; 18%, 5-9; 23%, 10-19; 18%, 20 or more). About 70% of the migraine patients consulted a general practitioner for their headaches. Only half of them did so during the past year and only 6% of them consulted a neurologist during that year.

TABLE 6.4.2 Respondents' characteristics in the migraine group (n=436) and control group (n=575)

	Migraine	Controls
Sex (% female)	84	80
Age [X, (sd)]	40 (13)	41 (14)
Employment status (% with paid job)	47	44
Education		
Low	34%	31%
Medium	38%	38%
High	28%	31%
Comorbidity (excl. migraine and headache)		
0 conditions	29%	43%
1 conditions	31%	27%
2 conditions	22%	14%
> 2 conditions	19%	16%
Number of conditions [X (sd)]	1.50 (1.54)	1.15 (1.40)

6.4.3 Health status

The results of the *SF-36* (see Table 6.4.3.1 and Figure 6.4.3.1) show statistically significant worse functioning for the migraine group in all eight domains.

The differences are small to medium-sized. The differences between migraine patients and controls are the largest for Pain, Social Functioning, Vitality and Role Limitations due to physical problems.

The NHP-1 results (see Table 6.4.3.2 and Figure 6.4.3.2) show significant results only for the scales Energy and Emotional reactions. The effect sizes are small. The results for the NHP-2 (see Table 6.4.3.3) show that migraine causes significant problems for household work, social life, home life and sex life; the largest effects are medium-sized (household work and home life).

The scores of the migraine group and the control group for the COOP/WONCA charts are shown in Table 6.4.3.4. The lower level of functioning of the migraine group is significant for two out of six items, viz. Daily Activities (small effect) and Overall Health (medium effect).

Table 6.4.3.5 shows the EuroQol classification scores. The scores of the migraine group are indicative of significantly worse health status of the migraine group for the items Usual Activities, Pain/Discomfort and Anxiety/Depression as well as for the valuation of own health. The effect sizes of these differences are small to medium.

TABLE 6.4.3.1 SF-36. Migraine-group (n=436) and control group (n=575)

	Migraine	Controls		
	X (sd)	X (sd)	MWU (p-values)	Effect size (d)**
Physical Functioning*	85 (19)	86 (21)	.006	.07
Social Functioning	76 (21)	85 (21)	<.001	.39***
Role limitations (physical)	63 (40)	77 (36)	<.001	.34***
Role limitations (emot.)	75 (38)	81 (34)	.007	.17***
Mental Health	72 (19)	77 (18)	<.001	.25
Vitality	62 (19)	68 (18)	<.001	.35
Pain	65 (22)	78 (22)	<.001	.57
General Health Perceptions	68 (20)	73 (18)	<.001	.29

* all scales: 0 = bad functioning, 100 = optimal functioning
** *d* = .2 : small effect; *d* = .5: medium effect; *d* = .8 : large effect
*** Because of non-normal or non-continuous distribution of the data of these scales, use of effect size *W* is generally more appropriate. However, computation of *W*s did not change the conclusions.

FIGURE 6.4.3.1 SF-36 scores. Migraine group (n=436) and control group (n=575)

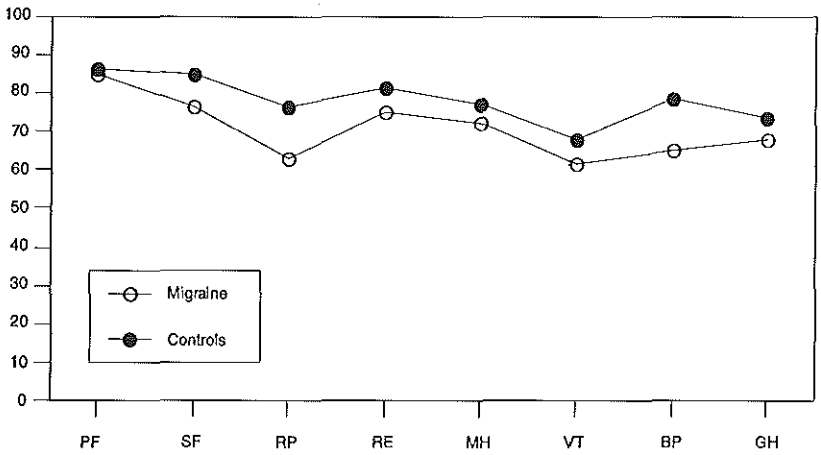


TABLE 6.4.3.2 Nottingham Health Profile(part 1). Migraine group (n=226) and control group (n=289)

	Migraine	Controls		
	X (sd)	X (sd)	MWU p-values	Effect size (*W*)**
Mobility*	9 (15)	6 (13)	.013	.12
Energy	20 (29)	12 (24)	.001	.15
Pain	11 (21)	7 (16)	.029	.12
Sleep	13 (24)	11 (22)	.221	.07
Social Isolation	8 (18)	5 (14)	.031	.10
Emotional Reaction	12 (18)	8 (17)	<.001	.21

* all scales: 0 = optimal level, 100 = worst level.
** *W* = .1 : small effect; *W* = .3: medium effect; *W* = .5 : large effect. *W* was used here instead of *d* because of non-normally or non-continuously distributed data.

FIGURE 6.4.3.2 NHP scores. Migraine group (n=226) and control group (n=289)

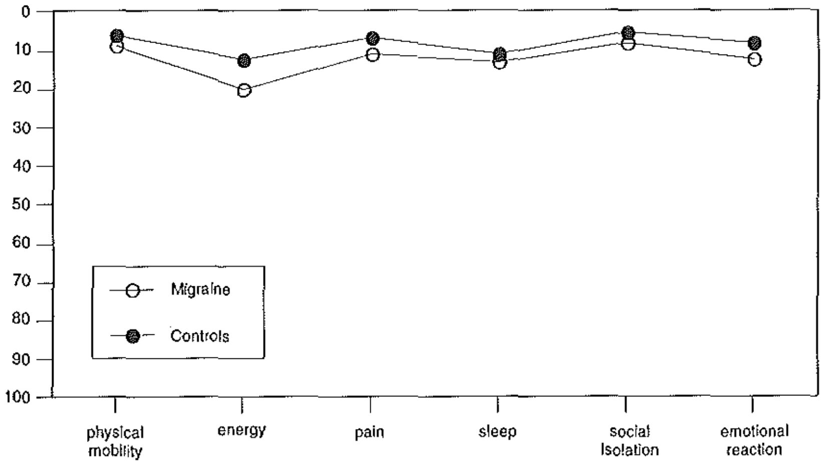


TABLE 6.4.3.3 Nottingham Health Profile (part 2). Migraine-group (n=226) and control group (n=289)				
	Migraine	Controls		
	(%yes)	(%yes)	Chi^2 (p-values)	Effect size (*W*)*
Health causes problems for ...				
...paid job	22	11	.34	.14
...household work	33	15	<.001	.21
...social life	25	11	<.001	.18
...home life	29	8	<.001	.27
...sex life	21	10	<.001	.15
...hobbies	22	14	.025	.10
...holidays	7	5	.335	.04

* *W* = .1 : small effect; *W* = .3: medium effect; *W* = .5 : large effect

6.4.4 *Consequences of comorbidity on functioning of migraine patients*

The results of the study as described above showed worse functioning of the migraine group *and* a higher prevalence of self-reported comorbid conditions, especially 'depression/nervous exhaustion' and 'diseases of the skin/eczema'. We examined the extent to which the impaired health status of the migraine sufferers could be attributed to migraine and to the most relevant comorbid conditions respectively. We did seven consecutive MCAs with Pain$_{(SF-36)}$, Role limitations (physical)$_{(SF-36)}$, Vitality$_{(SF-36)}$, Social Functioning$_{(SF-36)}$, General Health Perceptions$_{(SF-36)}$, Energy$_{(NHP)}$ and Valuation of own health$_{\text{(EuroQol)}}$ as dependent variables respectively.

Each of these MCAs showed significant coefficients for the explanatory variables 'migraine' and for 'depression' (p's<0.001), but insignificant coefficients for 'diseases of the skin'. The effect of 'depression' was larger than the effect of 'migraine', except for Pain$_{\text{(SF-36)}}$. For some of the dependent variables (Social Functioning$_{\text{(SF-36)}}$, Valuation of own health$_{\text{(EuroQol)}}$, Role limitations - physical$_{\text{(SF-36)}}$) the interaction effect (migraine* depression) was significant (p's<.01, .01 and .02 respectively), which indicates that the detrimental effect of the presence of both conditions on the dependent variable is larger than the additive effect of each of them.

TABLE 6.4.3.4 COOP/WONCA charts, migraine group (n=210) and control group (n=286)

	Migraine	Controls		
	X (sd)	X (sd)	MWU (p-value)	Effect size (*d*)**
Physical fitness*	1.7 (1.0)	1.7 (1.1)	.981	.00
Feelings	2.0 (1.0)	1.8 (0.9)	.031	.22
Daily activities	1.9 (0.9)	1.6 (0.8)	<.001	.29
Social activities	1.6 (0.8)	1.5 (0.8)	.210	.19
Change in health	2.6 (0.8)	2.7 (0.8)	.098	.10
Overall health	2.9 (0.9)	2.5 (1.0)	<.001	.39

* all charts: 1 = optimal level, 5 = worst level
** *d* = .2 : small effect; *d* = .5: medium effect; *d* = .8 : large effect

6.5 Conclusion and discussion

Our study shows that the health status of migraineurs is significantly impaired in comparison with a control group. The direction of the differences consistently indicated a worse health status of the migraineurs, regardless of the instrument used.

The fact that these differences were found with generic instruments, which are intended for assessment of health status ranging from 'very bad' to 'very good' is an indication that they are real differences. Because statistical significance is distinct from relevance, the differences between migraine and control group were placed in perspective by effect size estimators. The sizes of the differences were small to medium.

This finding has face-validity; despite the impaired functioning of migraineurs, migraine is generally not a severely incapacitating condition like, for example, end-stage cancer. Comparison of the results of the health status assessments of the migraine sufferers with published results for other patient groups are likely to be flawed to some extent because of different composition of the groups regarding, for example, age and sex. With this precaution in mind, the NHP scores of the migraine group in our study are in the same range as those of a group of Dutch patients with mild airflow obstruction.[30]

The largest differences between migraine sufferers and controls were observed in the domains Pain$_{(SF-36)}$, and to a lesser extent Pain/discomfort$_{(EuroQol)}$; Role limitations (physical)$_{(SF-36)}$, Household work$_{(NHP-2)}$; Social Functioning$_{(SF-36)}$, Homelife$_{(NHP-2)}$; Vitality $_{(SF-36)}$ and Energy$_{(NHP)}$, Overall health$_{(COOP)}$ and Valuation of own health$_{(EuroQol)}$. The unexpected lack of a difference on Pain$_{(NHP)}$ can be attributed to the fact that many of the items of this scale relate to pain when walking or standing; Pain$_{(SF-36)}$ refers more generally to the amount of bodily pain experienced in the past four weeks and its interference with normal work.

TABLE 6.4.3.5 EuroQol classification. Migraine group (n=436) and control group (n=575)

	Migraine	Controls		
			Chi^2	Effect size
	(%)	(%)	(p-values)	(W)**
Mobility*:				
no problems	83.0	87.4	.051	.06
some problems	16.3	12.4		
confined to bed	0.7	0.2		
Self-care:				
no problems	97.2	98.0	.386	.03
some problems	2.8	2.0		
unable to	0	0		
Usual activities:				
no problems	72.4	83.3	<.001	.15
some problems	26.4	14.8		
unable to	1.2	1.9		
Pain/Discomfort:				
none	49.5	67.4	<.001	.19
some	46.6	31.2		
extreme	3.9	1.4		
Anxiety/Depression:				
none	73.0	85.3	<.001	.15
some	24.6	13.7		
extreme	2.5	1.1		
Valuation of own health				
(0-100): X (sd)	77 (17)	83 (15)	<.001	.38 (*d****)

* 1=optimal level, 3=worst level
** *W* = .1 : small effect; *W* = .3: medium effect; *W* = .5 : large effect
*** *d* = .2 : small effect; *d* = .5: medium effect; *d* = .8 : large effect

Additionally we explored whether (self-reported) depressive disorders and diseases of the skin, which have a higher prevalence among migraineurs, could explain their health status impairment. The effect of migraine on health status remained significant after correction for these two conditions, which means that migraine has a consistent independent, though moderate, impact on health status.

The relevance of the presented results is twofold. Firstly, the impairment of the overall functioning of migraine patients has been quantitatively documented. Secondly, we showed the effect of migraine on health status to be independent of two relevant comorbid conditions, viz. self-reported depression and diseases of the skin. The impact of migraine on health status justifies the continuing search for cost-effective remedies for this condition. Treating migraine will probably improve the sufferer's functioning. However, migraineurs are at a greater risk of depression and other comorbid conditions, some of which have an additional detrimental effect on health status. Clinical awareness may result in a higher opportunity of treating these associated conditions, with probably additional positive effects on the daily functioning of migraine sufferers.

Acknowledgement

The authors wish to thank the respondents, Neil K Aaronson PhD, Arjan P Bandel BA, Bowine C Michel MD, Karien Stronks MSc, AM Eschauzier and Glaxo B.V. for their valuable contributions to the research underlying this paper.

References

1 Rasmussen BK, Breslau N. *Migraine - Epidemiology.* In: Olesen J, Tfelt-Hansen P, Welch KMA (eds). The Headaches. New York, Raven Press, 1993.

2 Royen L van, Essink-Bot ML, Koopmanschap MA, Michel BC, Rutten FFH. *A society's perspective on the burden of migraine in the Netherlands.* Pharmacoeconomics. 1995;7(2):170-179.

3 Headache Classification Committee of the International Headache Society. *Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain.* Cephalalgia 1988;8(Suppl 7):1-96.

4 Osterhaus JT, Townsend RJ. *The quality of life of migraineurs: a cross-sectional profile.* Cephalalgia 1991;11 (Suppl. 11).

5 Solomon GD, Skobieranda FG, Gragg LA. *Quality of life and well-being of headache patients: measurement by the Medical Outcomes Study Instrument.* Headache 1993;351-358.

6 Jenkinson C. *Health status and mood state in a migraine sample.* The International Journal of Social Psychiatry 1990;36:42-48.

7 Bos GAM van den. *Zorgen van en voor chronisch zieken (Concern of and for chronic patients).* Thesis. Utrecht, The Netherlands: Bohn, Scheltema & Holkema, 1989.

8 Breslau N, Davis GC, Andreski P. *Migraine, psychiatric disorders, and suicide attempts: an epidemiologic study of young adults.* Psychiatry Research 1991;37:11-23.

9 Jarman J, Fernandez M, Davies PT, Glover V, Steiner TJ, Thompson C, Rose FC, Sandler M. *High incidence of endogenous depression in migraine: confirmation by tyramine test.* J Neurol Neurosurg Psychiatr 1990;53:573-575.

10 Merikangas KR, Angst J, Isler H. *Migraine and psychopathology. Results of the Zurich cohort study of young adults.* Arch Gen Psychiatry 1990;47:849-853.

11 Mortimer MJ, Kay MB, Gawkrodger DJ, Jaron A, Barker DC. *The prevalence of headache and migraine in atopic children: an epidemiological study in general practice.* Headache 1993;427-431.

12 Chen TC, Leviton A. *Asthma and eczema in children born to women with migraine.* Arch Neurol 1990;47:1227-1230.

13 O'Keeffe ST, Tsapatsaris NP, Beetham WP. *Association between Raynaud's phenomenon and migraine in a random population of hospital employess.* J Rheumatol 1993;20:1187-1188.

14 Pal B, Gibson C, Passmore J, Griffiths ID, Dick WC. *A study of headaches and migraine in Sjögren's syndrome and other rheumatic disorders.* Ann Rheum Dis 1989;48:312-316.

15 Riera G, Vilardell M, Vaque J, Fonollosa V, Bermejo B. *Prevalence of Raynaud's phenomenon in a healthy Spanish population.* J Rheumatol 1993;20:66-69.

16 Passchier J, Andrasik F. *Migraine - psychological factors.* In: The headaches. J Olesen, P Tfelt-Hansen, KMA Welch (eds.). Raven Press Ltd, New York, 1993.

17 Ware JE. *Generic versus specific health status measures.* Medical Outcomes Trust Bulletin, 1994;2:4.

18 Ware JE, Sherbourne CD. *The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection.* Med Care 1992;30:473-483.

19 McHorney CA, Ware JE, Raczek AE. *The MOS 36-item Short Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs.* Med Care 1993;31:247-263.

20 McHorney CA, Ware JE, Lu JFR, Sherbourne CD. *The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of Data quality, scaling assumptions, and reliability across diverse patient groups.* Med Care 1994;32:40-66.

21 Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, Bungay K, et al. *International quality of life assessment (IQOLA) project.* Quality of Life Research 1992;1:349-351.

22 Hunt S, McEwen J, McKenna SP. *Measuring health status.* London, Croom Helm, 1986.

23 Bonsel GJ, Essink-Bot ML, Klompmaker IJ, Slooff MJH. *Assessment of the quality of life before and following liver transplantation.* Transplantation 1992;53:796-800.

24 Essink-Bot ML, Agt HME van, Bonsel GJ. *NHP or SIP: a comparative study in a group of chronically ill patients (in Dutch; English abstract).* Tijdsch Soc Gezondheidsz 1991;70:152-159.

25 Essink-Bot ML, Stouthard MEA, Bonsel GJ. *Generalizability of valuations on health states collected with the EuroQol questionnaire.* Health Economics 1993;2:237-246.

26 Weel C. van. *Functional status in primary care: COOP/WONCA charts.* Disability and Rehabilitation 1993;15:96-101.

27 Cohen J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1977.

28 Andrews FM, Morgan JN, Sonquist JA, Klem L. *Multiple classification analysis.* Institute for Social Research, The University of Michigan, Ann Arbor, Michigan, 1973.

29 Tabachnick BG, Fidell LS. *Using multivariate statistics.* Harper Collins Publishers, New York, 1989.

30 Schayk CP van, Rutten-van Mölken MPMH, Doorslaer EKA van, Folgering H, Weel C van. *Two-year bronchdilator treatment in patients with mild airflow obstruction.* Chest 1992;102:1384-1391.

7

Assessment of the quality of life before and following liver transplantation : first results

Published as:
Bonsel GJ, Essink-Bot ML, Klompmaker IJ, Slooff MJH. Assessment of the quality of life before and following liver transplantation: first results. Transplantation 1992;53:796-800. Reprinted with permission of the publisher.

7.1 Abstract

Analysis of the quality-of-life effects was part of the technology assessment of the Dutch orthotopic liver transplant program.

Data were collected by means of computer-assisted interviewing, including one interview before transplantation and annual follow-up interviews. Data on psychiatric morbidity were obtained from medical records.

This article shows preliminary results of a cross-sectional analysis of data collected from 1987 to 1989. Eighty-eight measurements were obtained from 46 adult patients (response rate 82%). Pre-transplant scores suggest major restrictions on all domains of life, especially a low amount of energy. After transplantation all indicators show improvement, although the level of the general population is not always attained. Improvement of subjective quality of life is more marked, probably due to euphoria at surviving the hazardous procedure. Psychiatric events occurred only infrequently.

We conclude that orthotopic liver transplantation contributes positively to the quality of life of surviving patients. Additionally, empirical health status assessment in these sometimes very ill patients appeared to be feasible.

7.2 Introduction

Though never assessed in a randomized clinical trial, orthotopic liver transplantation has been accepted as an effective therapy for patients with end-stage liver disease.[1] The considerable claim on financial resources and uncertainty about the balance of harm and benefit have given rise to many investigations of its overall effectiveness. Results on the gain in length of survival of patients treated with liver transplantation have recently become available.[2,3,4] However, studies assessing the gain in quality of life as a result of liver transplantation are few.[5,6] Most studies on the effect of liver transplantation on the quality of life have studied a few noncomprehensive indicators of quality of life, - e.g., days spent in hospital, work, activity status, and growth in height (children).[7,8,9,10,11,12,13] In addition several reports are available on psychiatric morbidity, focusing primarily on neuropsychiatric symptoms of end-stage liver disease (hepatic encephalopathy).[14,15,16]

A more comprehensive assessment of quality of life in liver transplantation patients was reported for liver transplant recipients only after the procedure, probably due to considerable difficulties in obtaining measurements from pretransplant patients. [17,18,19,20] The study by Lowe is of particular relevance since it applied a quality of life instrument (the Nottingham Health Profile) that was also used in our study.[20] Only the study published by Tarter in 1988 has a longitudinal design; preliminary results indicated a sharp improvement in quality of life after liver transplantation, although the premorbid level of quality of life usually was not attained.[21] The present article describes the design and the first results of a study on the changes in health-related quality of life in adult Dutch liver transplant patients. The study was part of a medical technology assessment of the liver transplant program in the Academisch Ziekenhuis Groningen in the Netherlands that analyzed effects of liver transplantation on survival, quality of life, and various resources.[3,22]

7.3 Materials and methods

7.3.1 *Patients*

The protocol of patient selection, timing of transplantation and support during follow-up has been described before.[23,24,25] The nationally accepted protocol states overt psychiatric morbidity, including active alcoholism, to be a contraindication. Intensive support by a specialized social worker is a regular part of the clinical program.

A longitudinal study design for the quality of life study with pre- and posttransplant measurements was aimed at. Until the start of data collection for the quality of life study in June 1987, 63 patients received a transplant; 38 of these were still alive from 1 month to 8 years posttransplantation. These 38 patients constituted the group that was only eligible for posttransplant measurements (the cross-sectional group), for obtaining information on the long-term quality of life. Eight patients under 18 were not elegible for the quality of life study; additionally 4 adults were left out for practical reasons as they lived abroad.

The remaining 26 Dutch adults living with a transplant in June 1987 were included in our study and completed a questionnaire once a year.

The longitudinal group consisted of all adult Dutch-speaking liver transplantation candidates who have entered the program since June 1987, as defined by the formal request for a donor liver from Eurotransplant. Collection of data for the present report was terminated on July 1st, 1989. By then the longitudinal group consisted of 26 patients, 6 of them still awaiting transplantation.

7.3.2 *Questionnaire*

It is generally recognized that health-related quality of life (or 'health status') constitutes a complex, multidimensional construct.[26] According to present standards objective and subjective components are discerned.[27] Objective quality of life usually refers to observable phenomena that can be compared with external standards (e.g., walking distance). Subjective quality of life refers to experienced well-being. We selected a set of general and specific questionnaires that addressed objective and subjective quality of life. If possible validated Dutch questionnaires were used. The left-hand columns of Table 7.3.2 show the questionnaires used as well as their ranges of scores and reference scores for the general population, if available.

The Nottingham Health Profile (NHP, part 1) is a comprehensive measure designed to measure perceived health on 6 specific domains of life, as shown in Table 7.3.2.[28] The NHP consists of 38 items with a yes/no answering format and was used by Lowe in his assessment of post-transplant status of liver transplantation patients.[20] The Karnofsky index is a global one-item measure for health status, often used in oncologic research.[29] It covers domains like intensity of treatment and ability to take care of oneself. The Index of Well-being is a global measure for experienced well-being, consisting of 11 items.[27]

The other questionnaires mentioned in Table 7.3.2 concern more specific indicators. The State-Trait Anxiety Inventory (STAI) and Self-rating Depression Scale (SDS or Zung), are 20-item questionnaires to measure anxiety and depression respectively.[30,31,32,33] The questions on Activities of Daily Life were derived from a Dutch national survey on health related problems. For nine activities, ranging from dressing to shopping, patients were asked whether they performed these activities independently - and, if they did, at what effort. The questions on physical complaints and working capacity were designed for this study. Inquiries were made about the following complaints by means of a three-point scale (absent/sometimes present/ always present): lack of appetite, abdominal cramps, swollen belly, itching, jaundice, bone pain, backache, hematomas, drowsiness. The questions on satisfaction with aspects of life originate from the Dutch health survey mentioned above.

The resulting questionnaire consisted of about 250 items. Stand-alone computer assisted interviewing was used as method of presentation of the questions and registration of the response.[34] This technique was succesfully applied earlier with ambulatory as well as bedridden patients in a similar study of heart transplantation patients.[35]

In addition to the self-reported quality of life, medical records of all patients were abstracted for the presence of psychiatric events. A psychiatric event was defined as clinical or outpatient treatment by a psychiatrist and/or the prescription of psychiatric drugs (excluding temporary prescription of benzodiazepine-derivatives).

TABLE 7.3.2 Quality of life before and after liver transplantation (reference values, mean patient values and standard deviations; 1987-1989, n=46, cross-sectional analysis)

	Reference values		Results in liver transplantation-population				
Questionnaire	Range**	General population	Waiting List (n=22)	3 months post LTx (n=18)	1 year post LTx (n=13)	2-5 yrs post LTx (n=15)	6-10 yrs post LTx (n=10)
General indicators:							
Nottingham Health Profile - 1							
Mobility	100-0	<15	34 (29)	24 (26)	13 (27)	15 (24)	9 (14)
Pain	100-0	<15	19 (25)	9 (18)	2 (5)	8 (17)	2 (4)
Energy	100-0	<15	63 (42)	18 (29)	3 (10)	11 (29)	9 (20)
Sleep	100-0	<15	42 (35)	14 (20)	9 (18)	5 (7)	10 (6)
Social isolation	100-0	<15	18 (23)	7 (16)	2 (5)	7 (13)	2 (6)
Emotional reaction	100-0	<15	14 (21)	7 (17)	3 (7)	5 (8)	5 (9)
Karnofsky-Index	0-100	≥ 90	64 (18)	71 (15)	87(13)	85 (16)	90 (5)
Index of Well-Being	2.1-14.7	> 12	9.5 (2.9)	13.5(1.1)	13.2(2.0)	13.4(2.3)	13.4(2.0)
Specific indicators:							
State-Trait Anxiety Inventory	80-20	≤ 37	41(11)	34(10)	29(7)	30(7)	32(5)
Self-rating Depression Scale (Zung)	100-25	≤ 33	50(11)	43(8)	39(5)	44(7)	43(6)
Activities of daily life	1-10	≥ 9	8.7(1.7)	9.2(1.4)	9.3(1.9)	9.4(1.4)	9.8(0.6)
Physical complaints	10-1	≤ 2	5.6(2.0)	3.4(1.9)	2.6(1.9)	3.4(1.8)	2.8(1.9)
Working* activity (hours/day)	0-12	± 8	1.9(3.0)	1.7(1.8)	3.3(2.3)	4.1(2.6)	5.0(2.5)
Median value			0.0	1.0	4.0	4.0	5.0
Satisfaction with							
...health	5-1	<2.6	4.1(1.0)	2.2(1.1)	1.6(0.8)	1.6(1.0)	1.4(0.7)
...leisure time	5-1	<2.6	3.2(1.6)	2.3(1.3)	1.7(0.8)	1.7(0.9)	1.5(0.7)
...daily activities	5-1	<2.6	3.1(1.5)	2.3(1.0)	1.7(0.9)	1.8(1.1)	2.0(1.0)
...life as a whole	5-1	<2.6	3.0(1.1)	2.2(1.2)	1.5(0.7)	1.5(0.8)	1.4(0.5)

* paid and unpaid work (e.g., housework, study); ** worst possible score on the left, best possible score on the right

7.4 Results

7.4.1 Response

Of 57 theoretically possible measurements among the 26 cross-sectional patients, 42 (75%) were actually realized. From all of them at least one measurement was obtained. Fourteen measurements were missed (including 2 refusals) from 12 patients due to initial organizational reasons unrelated to the physical condition of the patient. In the longitudinal group, 22 of 26 possible pretransplant measurements (85%) were obtained. Impaired physical condition of three patients prevented them from participation. Among the 20 patients meanwhile transplanted, five died. From the survivors, all 24 possible posttransplant measurements were obtained.

A total of 88 (42+22+24) measurements was obtained from 46 patients (overall response rate: 82%). The number of measurements related to the time of completion is presented in Table 7.4.1. As for paired observations (i.e., measurements obtained from the same patients before and after liver transplantation), 14 pre-3 months pairs could be obtained.

TABLE 7.4.1 Number of questionnaires completed

pre	3m	1yr	2yr	3yr	4yr	5yr	6yr	7yr	8yr	9yr	10yr	Total
22	18	13	8	5	3	4	5	4	4	1	1	88

7.4.2 *Patient characteristics*

Sociodemographic and medical characteristics at the time of transplantation are summarized in Table 7.4.2. Only one of the interviewed patients showed clear signs of hepatic encephalopathy preceding liver transplantation. She completed half of the questionnaire with considerable assistance. There was no significant difference in severity of disease, diagnosis, or other patient characteristics between earlier or later transplanted patients in the study-group; liver function of those alive at 1-year follow-up or more was good (data not shown).

7.4.3 *Quality of Life*

The interview results are summarized in Tables 7.3.2 and 7.4.3. The results shown in Table 7.3.2 represent a cross-sectional analysis. This implies that average results of measurements relate to groups of partially different composition. For reasons of presentation the results for the cross-sectional group of patients from 2 years after liver transplantation onward were combined into two groups: 2-5 years after liver transplantation and 6 years or more after liver transplantation (including only one measurement per patient per follow-up group). The results of the longitudinal analysis are shown in Table 7.4.3. Next we addressed the question of whether physical complaints were related to restrictions on particular domains of quality of life. The severity of the self-reported complaints were correlated with the dimension scores of the Nottingham Health Profile and with the Index of Well-being. Only Pearson correlations (r) exceeding 0.55 are mentioned. Restrictions on the mobility dimension of the Nottingham Health Profile correlate most with the presence of bone pain (r=0.57), on the pain dimension with the presence of bone pain (r=0.61) and backache (r=0.67), on the energy dimension with the presence of a swollen belly (r=0.60), on the sleep dimension with itching (r=0.57), on the emotional reaction dimension with drowsiness (r=0.59). None of the complaints correlated sufficiently with the score on the social isolation dimension of the Nottingham Health Profile. A low Index of Well-being correlated most with the presence of drowsiness (r=0.59).

Only three patients needed psychiatric treatment. One patient underwent psychiatric consultation preoperatively because of a suicide attempt in the past. One patient needed antipsychotic medication during the postoperative period. The third patient, who was actually transplanted for primary biliary cirrhosis, had a history of alcohol abuse. As anxiety to move hampered her remobilization, she was prescribed anxiolytic medication postoperatively.

TABLE 7.4.2 Sociodemographic characteristics of the study population (n=46)

	n	%
Sex		
female	31	67
male	15	33
Educational level		
minimal education	10	22
intermediate education	30	65
high education	6	13
Diagnosis		
cirrhosis excl. PBC	12	8
primary biliairy cirrhosis	17	37
primary sclerosing cholangitis	6	13
retransplantation	2	4
other	8	17
Child-Pugh classification at transplantation		
A	12	26
B	24	52
C	10	22
Age (years) at transplantation*		
X (sd)	42.6	(11.1)

* minus 5 patients not yet transplanted

TABLE 7.4.3 Quality of life before and 3 months after liver transplantation (n=14, longitudinal analysis, t-test)

Questionnaire	Waiting list	3 Months after LTx	p-value
General Indicators:			
Nottingham Health Profile-1			
Mobility	42	27	> 0.1
Pain	24	12	> 0.1
Energy	73	19	0.00
Sleep	47	16	0.01
Social isolation	17	8	> 0.1
Emotional reaction	23	8	0.07
Karnofsky Index	57	71	0.02
Index of Well-Being	8.9	13.3	0.00
Specific Indicators:			
State-Trait Anxiety Inventory	39	34	> 0.1
Self-rating Depression Scale	51	43	0.04
Activities of daily life	8.4	9.1	> 0.1
Physical complaints	6.0	3.4	0.01
Satisfaction with			
...health	4.4	2.3	0.00
...leisure time	3.7	2.4	0.02
...daily activities	3.4	2.3	0.04
...life as whole	3.4	2.1	0.01

7.5 Conclusions and discussion

In this article preliminary results of a study assessing the influence of liver transplantation on the quality of life are shown. The pretransplant state can be characterized by a rather low Karnofsky index, psychological distress, many physical disturbances and a low level of experienced well-being. Three months after transplantation patients show a considerable rise of the quality of life level. Further improvement in the first postoperative year results in a quality of life level similar to or slightly below the level of the general population. This level appears to stabilize in the following years. Evidently the very low frequency of long-term complications added to this favorable picture.

Experienced well-being shows an impressive favorable change following liver transplantation - to a level exceeding that of the general population. A comparison with the NHP-scores as reported by Lowe shows close similarity of the post transplant results, though comparison should be made with caution as the latter study combines data from patients shortly after transplantation (43% < 1 year) with data of patients at more than 2 years of follow-up.[20]

We are aware that our conclusions depend on some assumptions, especially with regard to the absence of a control group of nontransplanted patients and the effect of selective nonresponse. In this study, liver transplantation patients were used as their own controls. As gradual deterioriation of quality of life in the nontransplantation case is likely[36], a comparison of quality of life-values before and after liver transplantation may give a conservative estimate of the positive effect of liver transplantation.
Selective (non-) response is another threat to validity. High severity of disease precluded three patients, who can be expected to benefit most from liver transplantation, from completing the pretransplant questionnaire. Measurements from 5 posttransplant patients were missed as a result of mortality. The overall result of this selective nonresponse might be a slight overestimation of average pre- and posttransplant quality of life.

Some results deserve special attention. Firstly, the fairly good activities of daily life score preceding liver transplantation seems to contrast with the limited scores on work status and the Karnofsky index. Liver transplantation candidates are able to perform most daily activities independently only with great effort. Following liver transplantation these activities can be performed with little if any effort.
Secondly, the scores after liver transplantation on the SDS-Zung scale for depression seem rather high, - i.e., in the range of Dutch nondepressive psychiatric patients.[37] As reported by Smith, a high score does not necessarily mean the presence of a clinical depression following DSM-III standards.[38] Although the SDS-Zung scale is an accepted and validated questionnaire for depression, its interpretation for somatic patients remains to be established.
Thirdly, psychiatric morbidity occurred apparently only infrequently. The support by a specialized social worker probably contributes to this favourable outcome. Surman reported frequent episodes of pre- and postoperative anxiety and depressive disorders, and considered psychiatric consultation an èssential support to the transplant program.[16]
All together these results suggest that at present liver transplantation not only improves survival but also brings about significant improvement in quality of life. These findings should be validated by a longitudinal study, for which data are being collected now.

Acknowledgement

We wish to thank Mrs. C Steenmeyer, Mrs. M Messchendorp, and Dr. Els B Haagsma for their help as well as the National Health Insurance Board (Ziekenfondsraad) for financial support.

References

1 Neuberger JM, Adams DH. *Liver transplantation.* Bailliere's Clinical Gastroenterology 1989;3:231.

2 Bonsel GJ, Klompmaker IJ, Veer F van't, Habbema JDF, Slooff MJH. *Use of prognostic*

models for assessment of value of liver transplantation in primary biliary cirrhosis. Lancet 1990;i:493.

3 Bonsel GJ, Klompmaker IJ, Essink-Bot ML, Habbema JDF, Slooff MJH. *Cost-effectiveness analysis of the Dutch liver transplantation program.* Transplant Proc 1990;22:1481.

4 Starzl TE, Demeter AJ, Thiel D van. *Liver transplantation (first of two parts).* N Eng J Med 1989;321:1014.

5 U.S. Department of health and human services. *Assessment of liver transplantation.* Health Technology Assessment Reports, number 1. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, Department of Health and Human Services, 1990.

6 Pennington JC. *Quality of life following liver transplantation.* Transplant Proc 1989;21:3514.

7 Starzl TE, Koep LJ, Schroter GPJ, Hood J, Halgrimson CG, Porter KA. *The quality of life after liver transplantation.* Transplant Proc 1979;11:252.

8 Rolles K, Williams R, Neuberger J, Calne R. *The Cambridge and King's college hospital experience of liver transplantation, 1968-1983.* Hepatology 1984;4:50S.

9 Scharschmidt BF. *Human liver transplantation: Analysis of data on 540 patients from four centers.* Hepatology 1984;4:95S.

10 Williams JW, Vera S, Evans LS. *Socioeconomic aspects of hepatic transplantation.* Am J Gastroenterology 1987;82:1115.

11 Krom RAF, Wiesner RH, Rettke SR, et al. *The first 100 liver transplantations at the Mayo Clinic.* Mayo Clin Proc 1989;64:84.

12 Eid A, Steffen R, Porayko MK, et al. *Beyond 1 year after liver transplantation.* Mayo Clin Proc 1989;64:446.

13 Gartner JC, Zitelli BJ, Malatack JJ, Shaw BW, Iwatsuki S, Starzl TE. *Orthotopic liver transplantation in children: two-year experience with 47 patients.* Pediatrics 1984;74:140.

14 House R, Dubovsky SL, Penn I. *Psychiatric aspects of hepatic transplantation.* Transplantation 1983;36:146.

15 Tarter RE, Thiel DH van, Hegedus AM, Schade RR, Gavaler JS, Starzl TE. *Neuropsychiatric status after liver transplantation.* J Lab Clin Med 1984;103:776.

16 Surman OS, Dienstag JL, Cosimi AB, Chauncey S, Russell PS. *Psychosomatic aspects of liver transplantation.* Psychother Psychosom 1987;48:26.

17 Foley TC, Davis CP, Conway PA. *Liver transplant recipients - self-report of symptom frequency, symptom distress, quality of life.* Transplant Proc 1989;21:2417.

18 Wolcott D, Norquist G, Busuttil R. *Cognitive function and quality of life in adult liver transplant recipients.* Transplant Proc 1989;21:3536.

19 Roberts MS. *Quality-of-life measures in liver transplantation.* In: Quality of life and technology assessment. F. Mosteller and J. Falotico-Taylor, ed. Washington DC: National Academy Press, 1989.

20 Lowe D, O'Grady JG, McEwen J, Williams R. *Quality of life following liver transplantation: a preliminary report.* J R Coll Gen Pract 1990;24:43.

21 Tarter RE, Erb S, Biller PA, Switala JA, Thiel DH van. *The quality of life following liver transplantation: a preliminary report.* Gastroenterol Clin North Am 1988;17:207.

22 Bonsel GJ, Habbema JDF, Bot ML, Veer F van't, Charro FT de, Maas PJ van der. *A technology assessment of liver transplantation: a study of the Dutch liver transplantation programma in Groningen 1977-1987 (in Dutch).*Ned Tijdschr Geneeskd 1989;133:1406.

23 Putten ABMM van der, Bijleveld CMA, Slooff MJH, et al. *Selection criteria and decisions in 375 patients with liver disease, considered for liver transplantation during 1977-1985.* Liver 1987;7:84.

24 Slooff MJH, Bams JL, Sluiter WJ, et al. *A modified cannulation technique for veno-venous bypass during orthotopic liver transplantation.* Transpl Proc 1989;21:2328.

25 Klompmaker IJ, Haagsma EB, Gouw ASH, et al. *Azathioprine and prednisolone immunosuppression versus maintenance triple therapy including cyclosporine A for orthotopic liver transplantation.* Transplantation 1989;48:814.

26 Lohr KN. *Conceptual background and issues in quality of life.* In: Quality of life and technology assessment. F. Mosteller and J. Falotico-Taylor, ed. Washington DC: National Academy Press, 1989.

27 Campbell A, Converse PE, Rodgers WL. *The quality of American life.* New York: Russel Sage Foundation, 1976. ISBN 87154 194 7.

28 Hunt SM, McEwen J, McKenna SP. *Measuring health status.* London: Croom Helm; 1986.

29 Grieco A, Long CJ. *Investigation of the Karnofsky performance status as a measure of quality of life.* Health Psychol 1984;3:129.

30 Spielberger CD, Gorsuch RL, Lushene RE (eds). *STAI manual for the state-trait anxiety inventory.* Palo Alto, California: Consulting Psychologists Press Inc., 1970.

31 Ploeg HM van der. *Validation of the state - trait anxiety inventory, Dutch version (in Dutch).* Tijdschr Psychol 1980;35:243.

32 Zitman FG, Griez EJL, Hooijer Chr. *Standardization of depression questionnaires (in Dutch).* Tijdschr Psychiatr 1989;31:114.

33 Zung WWK, Durham NC. *A self-rating depression scale.* Archiv Gen Psych 1965;13:63.

34 Erdman HP, Klein MH, Griest JH. *Direct patient computer interviewing.* J Consult Clin Psychol 1985;53:760.

35 Bonsel GJ, Bot ML, Boterblom A, Veer F van't. *The costs and effects of heart transplantation. Volume 2C: Quality of life before and after heart transplantation, results (in Dutch).* Department of Public Health and Social Medicine, Erasmus University Rotterdam, The Netherlands, 1988. ISBN 90 72245 15 6.

36 Christensen E. *Individual therapy-dependent prognosis based on data from controlled clinical trials in chronic liver disease.* Dan Med Bull 1988;35:167.

37 Dijkstra P. *The self-rating depression scale of Zung (in Dutch).* In: Stemming en ontstemming. Praag HM van, Rooymans HGM, eds. Amsterdam: De Erven Bohn BV; 1974.

38 Smith MD, Hong BA, Robson AM. *Diagnosis of depression in patients with end-stage renal disease.* Am J Med 1985;79:160.

8

Evaluative health status measurement: an overview

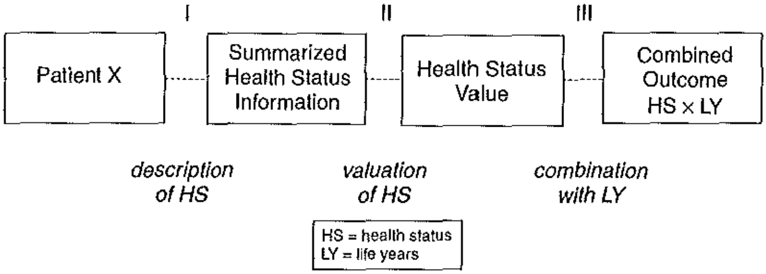
8.1 Introduction

As explained in chapter 2, application of health status data in assessing the burden of illness within populations, in public health models and in the comparative (economic) evaluation of medical interventions implies going one step beyond mere description of health status.[1,2,3] Descriptive instruments yield profile scores. If we are to judge whether one profile represents a better health status than another, i.e., if it is to be preferred over another, and if so, how much, a summary measure for score profiles is needed. Such summary figures ('values') may be obtained by means of a valuation procedure for health states. Once the values are available, they can be combined with survival data into a comprehensive outcome measure: for example, into quality adjusted life-years (QALYs) or disability adjusted life-years (DALYs).[4]
The present chapter provides an overview of the current scientific state of affairs in evaluative health status measurement. It concerns the valuation of health states. Many other important issues, such as the aggregation of individual outcomes to group outcomes and 'ethical' implications of QALYs are not addressed here.
In section 8.2 the empirical three-stage procedure that is currently followed in evaluative health status measurement will be explained. Important choices to be made at each stage will be elaborated in section 8.3. In section 8.4 we will explain some of the consequences of this three-stage approach, which is in fact an artificial disaggregation of the evaluation of composite outcomes characterized by quality and duration of survival. In section 8.5 we will demonstrate a tentative application of health status values in the evaluation of liver transplantation. Conclusions and recommendations for future research will be addressed in section 8.6.

8.2 Empirical evaluative health status measurement

The current three-stage approach to evaluative health status measurement will be introduced by an example.
Assume patient X suffers from a disease that causes physical and psychosocial dysfunction both to a moderate degree. Assume further that without treatment, patient X will live the following 2 years in the same health state. However, patient X might undergo an intervention that promises favourable effects with respect to both length and quality of life.

FIGURE 8.2 Current empirical three-stage approach to evaluative HSM



The effects of the intended intervention on the outcome of patient X will be evaluated. Figure 8.2 shows a diagram of the three-stage approach. In stage I, the descriptive stage, the health status of patient X is described, for example with a two-dimensional system consisting of a physical dimension A with three levels (1=best, 2=intermediate, 3=worst) and a psychosocial dimension B with three analogous levels. Let us assume that patient X's health status was assessed once before the intervention and twice afterwards, and that these assessments resulted in the following summarized health state descriptions:

- before the intervention: A_2B_2 (i.e., moderate physical and psychosocial functioning);
- 1 month after the intervention: A_2B_3 (i.e., moderate physical and bad psychosocial functioning);
- 6 months after the intervention: A_1B_2 (i.e., good physical, moderate psychosocial functioning).

So, the result of stage I is summarized information about the health status of patient X before and after the intervention. In the subsequent valuation stage (II), states A_2B_2, A_2B_3 and A_1B_2 should be valued. The health state descriptions should be in a type of format that enables their valuation. The subjects who perform the valuation task are requested to indicate, for example by means of a method called time trade-off (TTO), how good or how bad these three health states are, and to indicate their degree of undesirability. Assume the following values result from the value procedure: A_2B_2 0.70, A_2B_3 0.40, A_1B_2 0.80, where 1.00 = 'optimal death' ('most desirable') and 0.00 = 'death' ('most undesirable').

The values resulting from stage II are combined with life-years in stage III. If we assume that health status as measured at one month after the intervention is representative for the first period of three months after the intervention, and that the third assessment is representative for the remaining 21 months of the first two years after the intervention, we can compute the combined outcome of health status and life-years by simple linear weighting:

- without intervention: 2 years * .70 = 1.4 QALY;
- with intervention: (0.25 * .40) + (1.75 * .80) = 1.5 QALY.

The (hypothetical) intervention yields patient X a 0.1 QALY gain in the first two years after the intervention.

The above example is a simplification of the reality. Only one patient is evaluated. No mortality or other lasting undesirable effects are observed. To evaluate the effectiveness of an intervention, a well-designed randomized clinical trial evaluating groups of patients should be conducted. Both survival time and health status (until death) should be measured in the groups in both arms of the trial.

With respect to empirical health status assessment there are important issues to be decided on in stage I, i.e., the descriptive stage. These are:

a. *the domains of health status to be described and their operationalization, i.e., the choice of the dimensions;*

b. *methods to summarize the health state descriptions, so that they can be valued in stage II.*

Choices to be made in the design of the valuation study (stage II) relate to:

c. *the valuation method;*

d. *characteristics of presentation of the valuation task (stimulus presentation);*

e. *the subjects who perform the valuation task (the 'panel');*

f. *analysis and modelling of the panel data, leading to a comprehensive set of values for all health states than can possibly be described with the descriptive system chosen in b.*

After the assignment of panel-based values to each individual patient's health status descriptions in stage II, the final process which combines patients' data on health status and length of life into one figure involves one more choice, i.e.:

g. *whether or not to discount life-years or quality adjusted life years.*

Points a.-g. will be addressed in the next section.

8.3 Choices to be made in the current three-stage approach of outcome evaluation

a. *First issue of choice: the dimensions*

It was argued in Chapter 2 that the level of decision making at which information is to be used, or the perspective of the research question, determines the type of measure for descriptive health status assessment. If a societal viewpoint is adopted, the descriptive measure should be generic, i.e., comprehensive and non-disease specific. Generally, the choice of one of the 6 generic instruments addressed in Chapter 3 is equally defendable from a purely descriptive viewpoint. The subsequent valuation stage, however, has implications for the preceding descriptive stage, as will be explained in the next section.

b. *Second issue of choice: methods to summarize the health state descriptions*

Currently, health state values are often elicited by presenting a respondent with a description of a health state and asking him or her to rate how good or how bad that health state is according to his/her opinion. Health state descriptions are often presented as 'vignettes', see below.

TABLE 8.3.1 Example of a health state description (vignette): EuroQol state 21221

Some problems in walking about (2)
No problems with self-care (1)
Some problems in performing usual activities (2)
Moderate pain or discomfort (2)
Not anxious or depressed (1)

Descriptive measures for health status are available that provide a direct link to the valuation stage, e.g., the Quality of Well-being Scale (QWB), the Rosser-Kind Index and the EuroQol instrument.[5,6,7] A Dutch version is, as far as we know, only available for the latter instrument. These instruments either include a formalized procedure to translate health state descriptions into 'judgeable' vignettes (QWB, Rosser-Kind), or 'vignettes' are the direct result of the descriptive stage (EuroQol).

There are two options available if a descriptive measure is used that is not directly linked to health state valuations (for example: SF-36, MOS20/24, NHP, SIP, COOP/ WONCA charts). The first option is to use the existing link to values of other instruments: the researcher 'converts' a patient's SF-36 data (or from any of the other measures mentioned) into the dimensions of an instrument directly linked to health state valuations (e.g., EuroQol). This procedure was applied in the Dutch medical technology assessments of heart transplantation and liver transplantation.[8,9] The reliability and validity of the translation itself should be established.

The second option involves creating a direct link to health state valuations, i.e., the empirical valuation of health state descriptions resulting from the SF-36 or one of the other instruments. To allow for this option two conditions should be satisfied. Firstly, it must be possible to describe the scores on a scale by a system of mutually exclusive levels (in other words, a - probably ordinal - classification). Secondly, the number of dimension statements in the health state description should not exceed, say, 6 to prevent cognitive overloading of the respondents. Clearly, the creation of a direct link to health state valuations for the 5 generic instruments mentioned above is a relatively straightforward procedure only in the case of the COOP/WONCA charts (being a classification instrument itself).

c. Third issue of choice: the valuation method

Essentially two types of valuation methods are available. *Direct* valuation methods, e.g., a rating scale (RS), involve respondents being requested to attach a value to a given health state directly, i.e., to indicate the relative position of the state to be valued on a scale. *Indirect* methods, as represented by the standard gamble (SG) and time trade-off (TTO), involve deriving the value for a given health state from the amount of risk a subject is willing to take, or the number of life-years (s)he is willing to trade-off, respectively, to avoid being in the state to be valued. SG, TTO, and RS will be explained in more detail below. Alternative direct valuation methods, including magnitude estimation, and alternative indirect ones, including willingness-to-pay, will not be addressed here.

The standard gamble method essentially comprises an iterative paired comparison. The respondent is asked to select one alternative (s)he prefers from the two offered. His/her choice determines which alternatives will be offered in the iteration. SG is commonly operationalized as the choice between being in a specified lifelong stationary state (the state to be valued) on the one hand, and an intervention, for example a surgical procedure, with two possible outcomes, i.e., instantaneous and lasting improvement to perfect health (probability: p), or immediate death [probability (1-p)]. By varying p, the point of indifference between the two alternatives is determined. The more risk of immediate death the respondent is willing to take, the worse (s)he values the stationary state. SG is based on Von Neumann and Morgenstern's Theory of Expected Utility for decision-making under uncertainty.[10] SG includes two basic characteristics, i.e., choice and uncertainty.

Time trade-off was developed by Torrance as a less complicated but equally sound alternative for SG. The respondent is asked to trade-off length of survival and quality under conditions of certainty. The first alternative offers the respondent life in a (suboptimal) state for a fixed duration. The competing alternative offers a better health state (commonly optimal health) at a shorter duration. The indifference point is arrived at by varying the duration of optimal health. The shorter the period of optimal health the respondent is willing to accept, the lower (s)he values the alternative.

In methods which apply a rating scale, the respondent is offered a scale with labeled endpoints, for example 'death' and 'healthy', and asked to locate the state(s) to be valued on the scale. The EuroQol visual analogue scale, a thermometer with endpoints '0' (= the worst imaginable health state) and '100' (= the best imaginable health state) is an example of a rating scale (see Figure 10.3.3).

Several empirical studies aiming at comparing the results of different valuation methods have been reported; (see ref. 12, p.80 for a summarizing overview).[11,12,13,14,15] An important observation is that the ordinal ranking is generally not sensitive to the specific valuation method. A simple transformation from values obtained by one method into another is therefore theoretically possible.

With respect to numerical comparisons of valuations resulting from different valuation methods, there is empirical evidence suggesting that valuations resulting from SG and TTO procedures are to a large extent equivalent, whereas RS valuations are not equivalent to SG/TTO values. The relatively good health states especially are assigned lower values on a RS than in SG/TTO. This difference can intuitively be understood from the differences in the tasks. In SG and TTO respondents have to face a choice with important (hypothetic) consequences, while rating health states directly on a RS requires ranking without having to consider the consequences.

Although the results from the indirect methods are rather similar, empirical SG values are generally somewhat higher than TTO values. This observation is commonly explained by the fact that in SG risk-attitude plays a role. SG as a method to elicit values for health states is derived from game theory.[10] The behaviour of subjects in game-theoretical studies of individuals' preferences for goods such as money may be explained by a general risk-attitude and time-preference.

With respect to risk-attitude, a subject shows (by definition) risk-neutral behaviour if he is indifferent between a 50-50 gamble of obtaining 5 guilders or 15 guilders (expected utility of the gamble: 10 guilders) and a certainty of obtaining 10 guilders. If the subject prefers the gamble, (s)he is risk-seeking. If (s)he prefers the certain outcome, (s)he is risk-averse.

Time-preference refers to the timing of an outcome. A positive time preference indicates that an individual values goods higher if they are received now instead of later. Assume that the above example now offers a 50-50 gamble between obtaining 5 guilders or 15 guilders now, or a certainty of obtaining 10 guilders in one week's time. A subject with a positive time preference will prefer the gamble, a subject with a negative time preference prefers the certain outcome, whereas a subject without a time preference is indifferent. In this example, a positive time preference causes the subject to behave in a risk-seeking manner. Empirical evidence suggests that the risk attitude in SG is dependent on the context, i.e., the nature of the outcome (money or health).[16,17]

When applied to preference measurement for health states, Gafni and Torrance argued that preference behaviour in gambles involving outcomes consisting of combinations of duration (LY) and quality (Q) of survival can actually be disaggregated into three effects, i.e., a quantity effect, a time-effect and a gambling effect.[18] A gambling effect stands for a general fear of (or a liking for) gambles themselves. The effect may be dependent of the subject of the gamble. The quantity effect refers to the diminishing marginal value of additional LY*Q_x. The time effect represents a preference structure determined by the timing of the health gains. In SG the health states to be valued are presented for a given amount of life years, so that the quantity effect and the time effect are reduced to a single time preference effect.[17] This implies that a subject is risk-neutral in gambles with life-years if (s)he values each year of life equally *and* is neutral towards gambles.[17]

Miyamoto and Eraker operationalized a parameter r which is 'interpretable as a representation of a patient's risk attitude with respect to survival duration'.[19] They further state that:

> *'An individual is risk seeking if he always prefers a gamble with expectation Y to a certain survival of Y years. (..) For example, if faced with a choice between a certain survival of 6 years and a 50-50 gamble that yields either a 2-year survival or 10-year survival, a risk-averse individual would prefer the certain survival of 6 years, a risk-seeking individual would prefer the 50-50 gamble, and a risk-neutral individual would regard the certain option and the gamble to be equally desirable (or undesirable)' (ref. 19, p. 193).*

Following this operationalization, a risk-averse (r < 1.0) individual values years in the near future higher than years that are further away in the future, while a risk-seeking individual (r > 1.0) values later years more highly than years in the near future (ref. 19, fig 1). In other words, an individual that is risk-averse with respect to survival duration shows a positive time preference. Clearly, Miyamoto & Eraker's parameter r does not disentangle the gambling effect and the time preference effect, as was recognized by Stiggelbout et al.[13]

Miyamoto & Eraker proposed the method of certainty equivalents to estimate parameter r. If r < 1.0, indicating risk aversion with respect to gambles with life years, values elicited by SG are higher than those elicited by TTO. Adjustment of TTO values by r decreased the differences with SG both when the evaluators were students and patients.[12,13,19]

The observation that individuals may exhibit risk-averse or risk-seeking behaviour shows that they do not behave according to the axioms of expected utility theory. Kahneman & Tversky's Prospect theory may offer an explanation for the attitude towards risk in general; it argues that relative to a (hypothetical) individual reference point, outcomes are viewed as gains or losses, and risk attitudes vary depending on whether the outcome is seen as a gain or a loss.[20] A study by Verhoef et al. showed that the subjects were not risk-neutral with regard to life years.[21,22] Regret theory, suggesting that people may regret in the decisions they make, thus losing more utility than predicted by the expected utility approach, may offer an alternative explanation for general risk-attitude.[23]

d. Fourth issue of choice: stimulus presentation

Valuations of health states may be affected by the way the valuation task is presented to the raters, irrespective of the valuation method. The following heterogenous group of issues will be treated under this heading:

- Attributes of the 'vignettes': e.g., disease-labels, duration, prognosis.
- Alternatives to the 'vignette' mode of presentation of health states; for example, a multimedia presentation.
- Method of assessment: interview, paper-and-pencil, etc.
- Contextual effects: influence of presence of other states ('setting').
- 'Framing' effects.

Empirical evidence suggests that the addition of a *disease label* to the health state to be valued alters preferences.[24,25] We agree with the argument of Froberg & Kane that the addition of a disease label has the effect of providing more information for the subjects carrying out the valuation task.[26] Moreover, any labeling might convey negative connotations, e.g., about the duration and prognosis of a state, which might have an impact on the values. For example, it was shown that the use of the word 'cancer' negatively influenced subjects' valuations.[27]

Implicit or explicit statements about the *duration* of health states may be assumed to affect the values attached, even without taking into account the preceding state *(history)* and the state to follow *(prognosis)*. 'Health is unlike the usual outcomes studied in economics or decision theory. One reason is that the health of an individual has a time aspect inextricably bound to it.'[28] Sutherland et al. postulated the concept of 'maximal endurable time'.[29] They reported empirical evidence that attitudes of health professionals towards survival strongly depended both on the amount of time to be spent in a hypothetical health state and on the quality of the state. Subjects appeared to identify a personal variable, the maximal endurable time in a given state. When this time was exceeded, attitudes toward additional increments of survival changed dramatically. Recent data from York[30] supported this hypothesis: the longer a bad state lasts, the more intolerable it becomes and the lower the valuation assigned to that state. Duration and prognosis interact conceptually: a long duration of a bad state implies a bad prognosis, while a long duration of a good state implies a good prognosis. Kaplan et al. argued that prognosis of a state consists of the probabilities of transitions across function levels over time - in effect, the expected duration of a function level.[31] However, the authors treat the state to be valued, its duration and the following state as independent factors, i.e., without taking the possible interaction effects with respect to the valuation into account.[32]

Several studies have examined whether preferences shift due to the *mode of presentation* of the health states, for example written scenarios referring to laryngeal cancer patients compared with written scenarios plus a voice recording (Boyd 1982; cited in ref. 26); written outlines of scenarios compared with narrative style[33]; and a multimedia (computer screen + voice) presentation compared with narrative text.[34] In the latter study, the mean ratings for the two presentation modes were similar, but subjects who were presented with the multimedia description showed better processing of the information (indicated by better recall and better recognition). We consider that in general alternatives of 'vignette' presentation may inform the subjects performing the valuation task about the reality of the health state. Careful design of the presentation should prevent unintended subjective connotations similar to the situation as described above in relation to 'disease labels'.

Froberg and Kane cite one study showing that preferences were not significantly influenced by two *assessment methods*, i.e., the use of a computer compared with paper and pencil techniques.[35] There is some experience with methods other than interviewing in the context of the valuation of health states. One of the first studies with the EuroQol instrument [using a visual analogue scale (VAS)] dealt with the feasibility of collecting values elicited from a rating scale without outside assistance in a postal survey (see Chapter 9). The complexity of SG and TTO may preclude reliable and valid application without assistance. Collective evaluation of health states with SG and TTO with the help of slide presentation and a presenter in front of the floor appeared to be feasible in a group of students.[15]

Direct valuation methods commonly present the subject who performs the valuation task with a number of states simultaneously. For example, the EuroQol VAS presents 8 health states on one page of the valuation questionnaire, of which 4 are located on the left of a vertical VAS and the other 4 on the right, see Figure 10.3.3. We define *contextual effects* as the effects that the group of states as a part of which a state is presented (e.g., predominantly rather bad states, or predominantly rather good ones) might have on the rating assigned to that state. Such effects could not be shown in two experiments using category rating by Kaplan & Ernst (provided the endpoints of the scale are clearly defined)[36] nor in EuroQol VAS-data.[37]

Inconsistencies in valuations for health states which arise when the same objective alternatives are viewed in relation to different points of reference are called '*framing effects*'.[38] For example, respondents appear to show different preferences, depending on whether the certain outcome alternative in a gambling situation is presented as a gain or as a loss. In evaluative health status measurement, framing effects may occur (along with others) from differences in outcome descriptions and from a change in the anchoring (or reference) points.

With respect to outcome descriptions, McNeil et al. found that the attractiveness of surgery for lung cancer, relative to radiation therapy, was substantially greater when the information consisted of life expectancy rather than information regarding cumulative probability of survival, and when the problem was framed in terms of probability of living rather than in terms of the probability of dying.[39]

Empirical evidence suggests that both values elicited from rating scales[40] and values obtained with a standard gamble[41] alter with a change of the reference points. In the study by Sutherland, the ordinal ranking of the states was insensitive of different anchoring points. However, only a limited number (5) of states was valued. Values obtained by SGs in which one or both endpoints are replaced by any other health state can be rescaled so that they become comparable to values obtained with conventional SGs provided that the ranking of states remains unaffected, and provided that values for the alternative anchoring states are available.[14]

Froberg and Kane advocate helping the rater as much as possible and thus correcting the inconsistencies that occur as a result of limitations in human judgment, such as when the framing of a decision problem influences the rater's reference point.[26]

e. Fifth issue of choice: the subjects who perform the valuation task

It was argued in chapter 2 that the perspective from which a study is conducted determines conceptually whose values are the most appropriate to use. In studies from the patient group perspective the values of representatives of those patients are appropriate. If the health care policy perspective is adopted in a study, the values should represent the societal view. The latter is commonly operationalized by using the values of a representative sample (including patients and persons who have experiencé of bad health states) of the general population. If different groups in the general population, for example patients and healthy subjects, could be shown to hold different value patterns with respect to health states, the consequential question is how to aggregate thosė values to values reflecting the viewpoint of society.

With respect to the degree to which health status values are being affected by personal characteristics of the subjects who perform the valuation task Froberg & Kane conclude in their 1989 review[26] that the generally observed large inter-subject variation can only to a minor extent be explained by demographic characteristics (e.g., gender, age, education). The literature on rater differences suggests that only the rater's age and his/her experience of the health state being rated, respectively, may influence raters' valuations although even the evidence with respect to these variables is patchy.[42] For example, close agreement was found between weights for rating states as defined by the Quality of Well-being Scale obtained from the general population and from rheumatoid arthritic patients.[43]

It may be the case that that an effect of age on valuations of health states is in fact partly an effect of experience of suboptimal health states, because the probability of experience of states of illness increases with age. Previous experience of illness (affecting oneself, close relatives or friends, or subjects encountered professionally) probably makes the rater more aware of what bad states are actually like. Additionally, actually being in a state should be distinguished from past experience. Often patients *in a state* rate their own health state higher than healthy controls[44] or their proxies.[45] An interesting question is how such patients rate other states, both those that are better and worse than their own.[46]

We are not aware of any studies on the influence of personality characteristics, including, for example, neuroticism or coping style, on values of health states.

f. Sixth issue of choice: analysis and modelling of the value data elicited from a panel

The issue of aggregation of health state values from different respondents for one particular health state has been touched upon already in the preceding section. For 'common' data the choice of a measure of central tendency (mode, median, mean) depends on the level of measurement (nominal, ordinal, interval/ratio), and, to a lesser extent, the distribution of the data. In skewed distributions, the mean is sensitive for outliers, whereas the median is not. When aggregating the values attached to a health state by the members of a panel performing the valuation task the choice of a measure of central tendency is also a political one. If it is desirable to take values that deviate strongly from the majority into account, the mean is the appropriate measure of central tendency. If the majority rule of a voting procedure is applied, the median or even the mode would be more appropriate.

Ideally, empirical values should be available for all health state descriptions that can be constructed with a particular set of dimensions and categories. For example, the 5-dimensional, 3-level EuroQol allows for 243 (3^5) possible states. Due to the complexity and the cost of the task of obtaining empirical valuations, often only a limited number of these can be valued empirically. For example, the standard EuroQol valuation questionnaire, of which one page is shown in Figure 10.3.3, contains 13 states. This implies that a valid predictive model for the values for the empirically untouched states should be developed. Of course valid modelling needs empirical values for more than 13/243 = 5% of the states.

Often, empirical values are collected by using rating scales for reasons of feasibility (e.g., a rating scale being easier to understand and less time consuming than the indirect methods, thus allowing for a larger number of states to be valued per subject). Results from an as yet unpublished study by Krabbe et al also showed a VAS to be more reliable than either SG or TTO, at least among students. Several authors argue that TTO values are more valid than VAS values for use in economic outcome evaluation.[12,47,48,49] We think that in studies conducted from a societal viewpoint, the use of TTO values is more valid than the use of SG values, because at the *aggregate* level no uncertainty exists about the occurrence of different possible outcomes.

VAS values may be simply transformed into TTO values, provided that the ranking of the states is identical. Several authors suggest an exponential relation between VAS and TTO.[11,12]

$$\text{TTO} = 1\text{-}(1\text{-VAS})^{x}\,.$$

The value of the exponent x is 1.61 in Torrance's model, 2.12 in Busschbach's model and 2.32 in Krabbe's model (unpublished). The fact that Torrance's subjects performing the valuation task were selected from the general population, while Busschbach and Krabbe used students, may offer an explanation for the different values of x.

Theoretically, TTO values can be transformed to SG values, for example by empirical determination and application of Miyamoto & Eraker's exponent r (see section c.). If it is agreed that for decision-making at the aggregate level TTO values are more valid than SG values, transformation of TTO values to SG values may only be considered in supporting the decision-making process at an individual patient's level.

g. Seventh issue of choice: to discount or not to discount?

The steps described in the preceding sections result in a set of values for different health states. In the processing into QALYs, these values are combined with the durations of the states by simple linear weighting.

There is general agreement that in cost-effectiveness analysis costs should be discounted (i.e., future costs are valued as less important than costs to be made now). The rationale for discounting costs can be explained from the concept of time preference (see section c.) or from the concept of opportunity cost (if we have to spend money on a health care programme now we cannot spend the money on alternative activities). The discount rate is subject to debate. For reasons of comparability it is recommended to use a discount rate (r) of 5%, to perform sensitivity analysis and to always present results from the r = 0% variant. A detailed account on the discounting of future costs goes beyond the subjects of this thesis.

There has been considerable debate as to whether health effects should also be discounted. It has been argued that if costs are discounted, not discounting health benefits may lead to irrational or absurd consequences.[50] The time paradox of Cretin and Keeler implies that under the application of a lower (or zero) discount rate for health benefits than costs, every health programme can appear more cost-effective simply by delaying it.[51]

Discounting QALYs may result in double (dis)counting.[50] That is because, as addressed above in section c. (and see also section 8.4 below), in current methods of eliciting health state values (SG, TTO) time preference of individuals already plays a role. The magnitude of the error introduced by double discounting will depend on numerous factors, including the time frame used in utility assessment (the effect will be small if a short time frame is used).[50] The removal of the time preference effects from current valuation methods is not simple. Lipscomb suggested a method (called scenario strategy) enabling estimation of the magnitude of several effects (including time preference) on SG values, thus theoretically providing the opportunity to remove time preference effects from the quality weights preceding QALY calculations.[52] We are not aware of any application of this complicated method other than the study presented by Lipscomb that included valuation by 'undergraduate students taking a course in decision analysis'. Gafni argued recently that individuals' responses to time preference questions regarding health states are inevitably compounded by other effects that cannot be easily isolated, one of those being the sequence effect (see section 8.4).[28] Redelmeier et al. showed in a descriptive study that time preferences towards hypothetical health states vary widely among individuals, and that within individuals the discount rate is not constant over time and also depends on the type of health state.[53]

Despite the evidence that time preferences in individual choice behaviour cannot be described by a uniform discount rate, there are arguments to adhere to the current practice of discounting health benefits in cost-effectiveness analyses.[4,54,55] It is important to keep the perspective of a study in mind. If the study is conducted from a societal perspective, the use of average individual discount rates for health states (often estimated to lie between 2% and 10%) is appropriate and sensitivity analysis should determine how important the discount rate is to the results. In estimating the Global Burden of Disease a low positive discount rate of 3% was chosen for the calculation of Disability Adjusted Life Years.[4] A study from a patient's perspective requires an explicit elucidation of time preferences of the patients involved.[56]

8.4 The consequences of disaggregation of the outcome tree

In the preceding sections the current three-stage approach to evaluating outcomes has been explained. Outcomes are considered to be characterizable by duration and quality of survival, operationalized as life years and health status. In the current operationalization of the QALY or QALY type concept, health states are separated from their duration in the valuation stage. This procedure is actually a disaggregation of reality, as will be illustrated below. After that, some of the consequences of the disaggregation will be addressed.

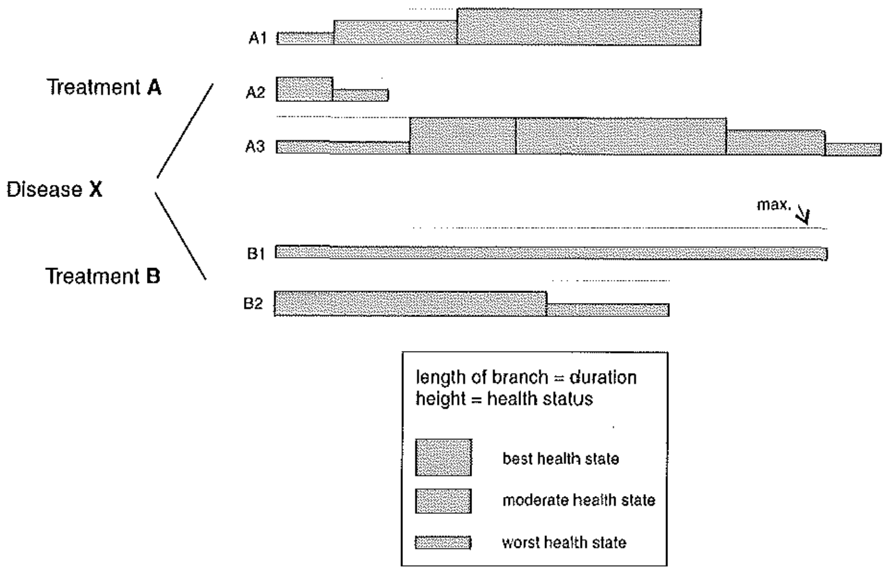
Comprehensive outcome measurement implies description and evaluation of the complete 'outcome space'. The outcome-space of a disease or of an intervention consists of multiple branches, each of them in the most complicated (often occurring) case consisting of a different number of life-years characterized by different health states occurring in different sequences, see Figure 8.4.1. Each branch is further characterized by a probability, which is on the aggregate level reflected by a frequency of occurrence.

When comparing the effects of diseases or interventions, essentially the complete outcome tree of each treatment option should be valued. If data scarcity did not prevent us from doing so, cognitive overloading of respondents would.

Consequentially, in the current operationalization of the QALY concept the outcome space was disaggregated for practical reasons. Health status is regarded separately from its duration in the valuation stage. Games such as SG and TTO are played in an attempt to obtain 'timeless' health status values. Subsequently these health status values meet their duration again in a procedure of combining life years and health status effects in each separate branch of the outcome tree. Then, quality adjusted life-years are discounted. The value of the outcome space is a weighted average of the number of discounted QALYs per branch. Some of the problems arising from this disaggregated approach are discussed below.

Firstly, a procedure where the quality of survival is valued without taking the duration into account assumes *mutual utility independence of life-years and health status*. This implies among other things that the duration of a health state should not affect the value of the health state, which assumption is generally not satisfied, for example if the existence of a 'maximal endurable time' concept is replicated (see section 8.3.d).

FIGURE 8.4.1 Example of an outcome-tree



Secondly, implicit *time preferences* continue to affect the values resulting from TTO and SG in spite of attempts to obtain timeless values. The TTO approach seems attractive, because quality is traded-off against duration and the resulting value is dimensionless. However, the time-unit used to elicit TTO values can be expected to affect the results. If a subject is requested, for example, to choose between 10 days in a state (followed by death) and less than 10 days in perfect health, (s)he will be likely to choose the ten days unless the state to be valued is so bad that a length of 10 days exceeds its 'maximal endurable time'. TTO in days leads to an unwillingness to trade-off any length of remaining life. This situation is similar to SG if the gamble includes a risk of immediate death, and subjects are thus asked to risk the loss of life-years in the very near future.

TTO tasks are often phrased in terms of an age-adjusted life expectancy; e.g., subjects in their thirties are presented with a life expectancy of, e.g., 40 years, whereas subjects aged 60 are confronted with, e.g., 15 years. The problem with this approach is that the younger subjects can be expected to trade-off more readily those far-away years at the expected end of their lives in order to avoid less-than-optimal health now. This problem is not solved by offering younger and older subjects similar time periods of, say, 10 years, as in that case younger subjects may be expected to be unwilling to trade-off any of those ten years because they want to live at least 10 years (to a large extent irrespective of the quality of those years), because, for example, they have young children to raise. Stiggelbout et al suggested obtaining information about the effect of time preference on TTO scores by repeating the TTO for various periods of time.[13] Johanneson et al. suggested deriving the QALY weights from TTO by dividing the number of discounted life years in full health by the number of discounted life years in the assessed health states.[57] In any case, the validity of the assumption of constant proportional trade-off (i.e., the loss of 1 year to a 5-year life span is equivalent to a loss of 10 years from a 50-years lifespan) may be doubted, as was also indicated by Kiebert et al.[16]

Direct rating by means of a rating scale also requires a statement of length. In the EuroQol VAS valuation questionnaire, respondents are instructed to imagine that the duration of each state to be valued is one year, while what happens afterwards is not known and should not be taken into account. Empirical evidence suggested that a majority of a group of 100 students performing the valuation task did not use this information in the subsequent valuation task.[12,58] However, the values resulting from the EuroQol questionnaire essentially hold for a duration of one year.

Thirdly, preferences for *sequences* of outcomes are not taken into account by the disaggregated operationalization of the QALY concept. A health state is valued in isolation. The validity of the assumption that the value of a state is independent of history and prognosis may be doubted. For example, a bad state may be tolerable for one day with the expectation of complete recovery and a better health state afterwards, whereas the same state lasting two months and a prognosis of gradual deterioration may be valued as worse than death.

The type of problems expected to occur when separate health states are valued instead of sequences will be illustrated by an example (see Figure 8.4.2).

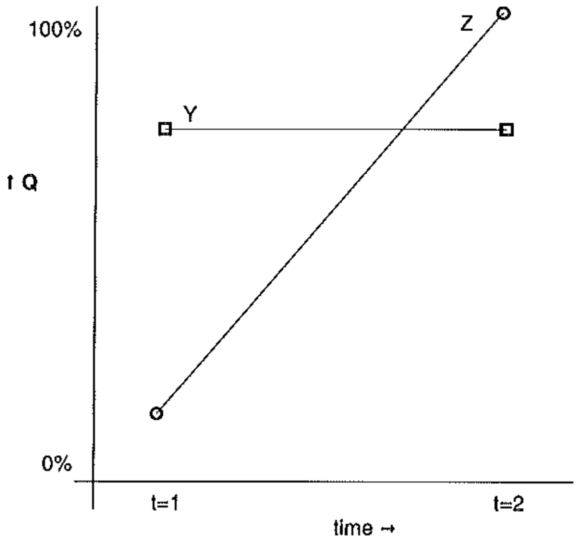
Assume patient Y with health status A_1B_2 and patient Z with health status A_2B_3 (the two-dimensional instrument is the same as in the example in section 8.2). The preference can be established easily: Y's health status is dominant over Z's.

'Assume Y's and Z's health status to develop over time. At a second assessment, Y's health status is still A_1B_2, but Z's is A_1B_1. At the second assessment, Z's health status is dominant over Y's. However, if we look at the course of health status, Y is in a stationary state, while Z has improved. Although it is easy to choose whose health status is better at each point of assessment, a preference for one of the scenarios is not straightforward. Evidently, more assessments over time may complicate the comparison even further.

Apart from the fact that sequences of outcomes are disregarded, the (dis)utility of a change is not valued either: changes are assumed to occur instantaneously and without burden.

The solution proposed by Mehrez & Gafni to overcome some of the problems addressed in this section (i.e., the Healthy Years Equivalents) has recently been shown to suffer from similar deficits as the conventional QALY approach.[49,59,60]

FIGURE 8.4.2 Different sequences of outcomes



8.5 A tentative application of health status values in the MTA of liver transplantation

The health status effects of liver transplantation can be demonstrated by the application of health status values. The data presented result from the study as described in chapter 7, supplemented with data from the extension of the project until 1991.[9,61]
In the first stage actual patients' health states, occurring before and at regular intervals after liver transplantation, were described using, among other instruments, the Nottingham Health Profile, see Table 8.5.1 [NHP scores range from 0 (= best possible score) to 100; general population norms < 15 for all dimensions].

TABLE 8.5.1 Nottingham Health Profile scores [mean (s.d.)] before and after liver transplantation (LTx)

NHP	Waiting list n = 43	3 months post LTx n = 31	1 year post LTX n = 25
Mobility	36 (29)	27 (25)	15 (27)
Pain	19 (27)	12 (19)	5 (20)
Energy	68 (40)	21 (30)	5 (15)
Sleep	43 (36)	17 (23)	14 (22)
Social Isolation	19 (23)	8 (16)	9 (21)
Emotional Reaction	21 (23)	8 (17)	4 (9)

After LTx, the health status of patients as measured by the NHP showed an improvement. The question is how much improvement? A second problem is that these 6-dimensional profile scores cannot easily be combined with life years into one outcome measure. Therefore, a measure is needed that summarizes each profile score into one figure. These summary figures should reflect the values of the respective health states from a societal viewpoint, as the liver transplantation study was a comprehensive MTA.

In the liver transplantation case, the procedure consisted of the following steps:

1. Recoding of the patients' NHP-scores at relevant assessment points into descriptions according to the dimensions of the EuroQol operationalization of health status.
2. Linking these health state descriptions to health state valuations from a sample of the general population, as collected using the standard EuroQol visual analogue scale (VAS).
3. Transformation of VAS values to TTO values: TTO = $1\text{-}(1\text{-VAS})^x$.
4. Sensitivity analysis for values of x.
5. Combination with life years gained into QALYs.

Steps 1. to 5. are illustrated below.

1. *Recoding NHP into EuroQol.* The recoding of NHP-scores into descriptive EuroQol scores was necessary here because the patients did not complete the EuroQol descriptive instrument themselves. The prototype of the EuroQol instrument became available only after the start of the datacollection in the liver transplantation study. In present studies, this recoding step is not required. The EuroQol system of dimensions of health status (see Table 10.3.2) consists of 5 dimensions (labeled Mobility, Self-care, Usual Activities, Pain/discomfort and Mood, respectively) with 3 ordered categories (1=no problems, 2=some problems, 3=extreme problems/unable to) each. Theoretically, 3^5 (=243) health state descriptions are possible.

   The patients' responses to the NHP were used to recode each patient's health state at the relevant assessment points into the EuroQol dimensions.
2. *Linking descriptions to VAS values.* Empirical values from a general population sample are available for 25 of the 243 EuroQol health states. These values were obtained by using the standard EuroQol Visual Analogue Scale (VAS) in a postal survey, (see Chapter 10). The endpoints of this VAS are 0 (= the worst imaginable health state) and 100 (= the best imaginable health state). Values for the remaining states were tentatively modelled.[62]
3. *Transformation of VAS values to TTO values.* VAS values are relatively easy to obtain. Time trade-off (TTO) is more difficult to operationalize, but several authors argue that TTO values are more valid for use in QALY calculations.[12,47,48,49]
4. *Sensitivity analysis for values of the exponent x.* Varying the exponent x may be regarded as a sensitivity analysis of the results for the valuation method employed. It is also necessary because the 'right' value of x has not been established (see section 8.3.f).

Table 8.5.2 shows the median values before (in the column x = 1.00) and after the exponential transformation (for a number of levels of x) of the health states observed before and at three moments after LTx.

TABLE 8.5.2 Median transformed [TTO = 1-(1-VAS)^x] values for health states observed before and after LTx

	x=.80	x=1.00	x=1.25	x=1.50	x=2.00	x=2.25	x=2.50
Waiting list	.43	.51	.59	.66	.76	.80	.83
3 months post LTx	.56	.64	.72	.78	.87	.90	.92
1 year post LTx	.69	.77	.84	.89	.95	.96	.97

From the results presented in Table 8.5.2 we may conclude that there is an important health status improvement after LTx (from waiting list to 1 year after LTx), the size of which may be estimated in a range between 0.14 - 0.26 on a 0-to-1 scale.

5. *Combination of health status values with life years.* Values for health states can be combined with the duration of those states. In the MTA of liver transplantation, the mean number of life years gained was estimated to be 3.8 (study horizon 10 years, 5% discount rate) and 7.6 (study horizon 25 years, 5% discount rate). When these life years gained were combined with estimates for the values for the health states by using VAS values after an exponential transformation as described above (using Torrance's value of the exponent), the mean numbers of QALYs gained were 3.5 and 6.9, respectively.[9,61]

8.6 Conclusions and recommendations

In this chapter we explained the current three-stage operationalization of the QALY concept, consisting of descriptive health status assessment, followed by a second stage in which health states are valued and a third stage consisting of combining life years with these quality estimates. We also showed some of the consequences of the disaggregation of the outcome tree, in which length and quality of survival are regarded separately during the valuation stage. However, despite its limitations, a demonstrable superior alternative to the existing operationalization of the QALY concept has still to be developed.

'But please do not be discouraged. Our experience is that in practice these measurements are not as onerous as they may at first appear. And our conviction is that for quality economic appraisals these measurements are often essential - for it is far better to have an approximate measure of the right factors than a precise measure of the wrong ones.'[3]

Further research is urgently needed especially on the following topics:

1. The role of 'time' in evaluative health status assessment. The problems occurring in the current approach of eliciting 'timeless' values of health states and the subsequent recombination with life-years can not easily be solved. In the valuation of outcome, length of survival, its quality, and the sequence of states are essentially inseparable. Ceasing to attempt to obtain timeless values, as they are essentially non-existent, might be a step in the right direction.
2. Population subgroups showing systematically different valuation patterns. If such subgroups cannot be shown to exist, the aggregation problem mentioned above in section 8.3.e. does not exist. We think that 'experience of bad health states' (past and/or present; in self, in close relatives/friends or professionally) is the only factor which might have a relevant effect on health state values. If population subgroups are shown to hold different value patterns with respect to health states, the consequential question is how to aggregate these values.
3. Modelling the valuation space. The claim of a 'direct' link of EuroQol health state descriptions to values should be interpreted with caution. Only a limited number of the theoretically possible states have been valued empirically, while adequate modelling of values of the remaining states is still a technical challenge.

References

1 World Bank. *World Development Report 1993: Investing in health - world development indicators.* Oxford: Oxford University Press, 1993.

2 Ruwaard D, Kramers PGN, eds. *Volksgezondheid Toekomst Verkenning.* Den Haag: SDU, 1993.

3 Drummond MF, Stoddart G, Torrance GW. *Methods for the economic evaluation of health care programs.* Oxford University Press 1987.

4 Murray CJL. *Quantifying the burden of disease: the technical basis for disability-adjusted life-years.* WHO Bulletin 1994;72(3):429-445.

5 Kaplan RM, Anderson JP, Ganiats TG. *The Quality of Well-being Scale: rationale for a single quality of life index.* In: Walker SR, Rosser R, eds. Quality of life assessment in the 1990s. Dordrecht: Kluwer Academic Publishers, 1993.

6 Kind P, Rosser RM. *The quantification of health.* Eur J Social Psychol 1988;18:63-77.

7 EuroQol Group. *EuroQol - a new facility for the measurement of health-related quality of life.* Health Policy 1990;16:199-208.

8 Bonsel GJ, Bot ML, Boterblom A, Veer F van 't. *Costs and effects of heart transplantation. Volume 2c: quality of life before and after hearttransplantation - results.* (De kosten en effecten van harttransplantatie. Deelrapport 2c: kwaliteit van leven voor en na harttransplantatie - resultaten, in Dutch) Rotterdam: Inst. Maatschappelijke Gezondheidszorg, 1988.

9 Michel BC, Bonsel GJ, Stouthard MEA, Essink-Bot ML, McDonnel J, Habbema JDF. *Liver*

transplantation: long-term effectiveness. (Levertransplantatie: effectiviteit op lange termijn, in Dutch). MGZ-report 92.07. Rotterdam: Department of Public Health, 1992.

10 Neumann J von, Morgenstern O. *Theory of games and economic behavior.* Princeton: Princeton University Press, 1953.

11 Torrance GW. *Social preferences for health states: an empirical evaluation of three measurement techniques.* Socio-Econ Plan Sci 1976;10:129-136.

12 Busschbach J van. *The validity of QALYs.* (De validiteit van QALY's, in Dutch). Thesis Erasmus University Rotterdam. Arnhem: Gouda Quint, 1993.

13 Stiggelbout AM, Kiebert GM, Kievit J, Leer JWH, Stoter G, Haes JCJM de. *Utility assessment in cancer patients: adjustment of Time Tradeoff scores for the utility of life years and comparison with Standard Gamble scores.* Med Decis Making 1994;14:82-90.

14 Rutten-van Mölken MPMH, Bakker CH, VanDoorslaer EKA, VanderLinden Sj. *Methodological issues of patient utility measurement: experience from two clinical trials.* Accepted for publication in Med Care.

15 Krabbe PFM, Essink-Bot ML, Bonsel GJ. *On the equivalence of collectively and individually collected response: standard gamble and time trade-off judgements of health states.* Submitted.

16 Kiebert GM, Stiggelbout AM, Kievit J, Leer JWH, Stoter G, Haes JCJM de. *Standard Gamble and Time Trade-off methods: are they in accordance with two basic assumptions of the QALY model?* Chapter 5 in: Kiebert GM. Choices in oncology: patients' valuations of treatment outcomes in terms of quality and length of life. Thesis. Leiden, 1995.

17 Hellinger FJ. *Expected utility theory and risky choices with health outcomes.* Med Care 1989;27:273-279.

18 Gafni A, Torrance G. *Risk attitude and time preference in health.* Management Science 1984;30:440-451.

19 Miyamoto JM, Eraker SA. *Parameter estimates for a QALY utility model.* Med Decis Making 1985;5:191-213.

20 Kahneman D, Tversky A. *Prospect theory: an analysis of decision under risk.* Econometrica 1979;47:263-291.

21 Verhoef LCG, Haan AFJ de, Daal WAJ van. *Risk attitude in gambles with years of life: empirical support for prospect theory.* Med Decis Making 1994;14:194-200.

22 Nease RF. *Risk attitudes in gambles involving length of life: aspirations, variations, and ruminations (Commentary).* Med Decis Making 1994;14:201-203.

23 Loomes G, Sugden R. *Some implications of the more general form of regret theory.* J Economic Theory 1987;41:270.

24 McNeil BJ, Pauker SG, Sox HC, Tversky A. *On the elicitation of preferences for alternative*

therapies. N Engl J Med 1982;306:1259-1262.

25 Sackett DL, Torrance GW. *The utility of different health states as perceived by the general public.* J Chron Dis 1978;7:347-358.

26 Froberg DG, Kane RL. *Methodology for measuring health-state preferences I - IV.* J Clin Epid 1989;42:345-354; 459-471; 585-592; 675-685.

27 Gerard K, Dobson M, Hall J. *Framing and labelling effects in health descriptions: quality adjusted life years for treatment of breast cancer.* J Clin epidemiol 1993;46:77-84.

28 Gafni A. *Time in health: can we measure individuals' pure time preferences?* Med Decis Making 1995;15:31-37.

29 Sutherland HJ, Llewellyn-Thomas H, Boyd NF, Till JE. *Attitudes toward quality of survival - the concept of 'maximal endurable time'.* Med Decis Making 1982;2:299-309.

30 Dolan P. *Duration study.* Paper presented at the sixth plenary EuroQol meeting, London, October 1994.

31 Blischke WR, Bush JW, Kaplan RM. *Successive intervals analysis of preference measures in a health status index.* Health Services Research 1975;181-198.

32 Kaplan RM, Bush JW, Berry CC. *Health status index. Category rating versus magnitude estimation for measuring levels of well-being.* Med Care 1979;XVII:501-525.

33 Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. *Methodologic issues in obtaining values for health states.* Med Care 1984;22:543-552.

34 Goldstein MK, Clarke AE, Michelson D, Garber AM, Bergen MR, Lenert LA. *Developing and testing a multimedia presentation of a health-state description.* Med Decis Making 1994;14:336-344.

35 O'Connor AM, Boyd NF, Till JE. *Influence of elicitation technique, position order and test-retest error on preferences for alternative cancer drug therapy.* Cited in: Froberger & Kane III, ref 26.

36 Kaplan RM, Ernst JA. *Do category rating scales produce biased preference weights for a health index?* Med Care 1983;XXI:193-207.

37 Stouthard MEA, Essink-Bot ML. *EuroQol 1991 - the Rotterdam survey - Results.* In: EuroQol Conference Proceedings. S. Björk, editor. Lund (Sweden): Institute for Health Economics, 1992.

38 Tversky A, Kahneman D. *The framing of decisions and the psychology of choice.* Science 1981;211:453-458.

39 McNeil BJ, Pauker SG, Sox HC, Tversky A. *On the elicitation of preferences for alternative therapies.* NEMJ 1982; 306:1259-1262.

40 Sutherland HJ, Dunn V, Boyd NF. *Measurement of values for states of health with linear*

analog scales. Med Decis Making 1983;3:477-487.

41 Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. *The measurement of patients' values in medicine.* Med Decis Making 1982;2:494-462.

42 Kind P, Dolan P. *The effect of past and present illness experience on the valuations of health states.* Med Care 1995;33:AS255-AS263.

43 Balaban DJ, Sagi PC, Goldfarb NI, Nettler S. *Weights for scoring the Quality of Well-being Instrument among rheumatoid arthritis.* Med Care 1986;24:973-980.

44 Stensman R. *Severely mobility-disabled people assess the quality of their lives.* Scand J Rehab Med 1985;17:87-89.

45 Tsevat J, Cook EF, Green ML. *Health status values of the seriously ill.* Ann Intern Med 1994;122:514-520.

46 Selai C, Rosser R. *Eliciting EuroQol descriptive data and utility scale values from inpatients.* PharmacoEconomics 1995;8:147-158.

47 Fanshel S, Bush JW. *A health status index and its application to health services outcomes.* Operations Research 1970;18:1021-1066.

48 Richardson J. *Cost utility analyses: what should be measured?* Soc Sci Med 1994;39(1):7-21.

49 Bonsel GJ. *Methods of medical technology assessment with an application to liver transplantation.* Thesis. Rotterdam, 1991.

50 Krahn M, Gafni A. *Discounting in the economic evaluation of health care interventions.* Medical Care 1993;31:403-418.

51 Keeler EB, Cretin S. *Discounting of life saving and other non-monetary benefits.* Management Science 1983;29:300.

52 Lipscomb J. *Time preference for health in cost-effectiveness analysis.* Med Care 1989;27:S233-253.

53 Redelmeijer DA, Heller MD. *Time preference in medical decision making and cost-effectiveness research.* Med Decis Making 1993;13:212-217.

54 Weinstein MC. *Time-preference studies in the health care context (Commentary).* Med Decis Making 1993;13:218-219.

55 Redelmeier DA, Heller DN, Weinstein MC. *Time preference in medical economics: science or religion? (Reply to commentary).* Med Decis Making 1994;14:301-303.

56 Ganiats TG. *Discounting in cost-effectiveness research (Commentary).* Med Decis Making 1994;14:298-300.

57 Johanneson M, Pliskin JS, Weinstein MC. *A note on QALYs, time trade-off and discounting.*

Med Decis Making 1994;14:188-193.

58 Busschbach JJ van, Hessing DJ, Charro FT de. *Observations on one hundred students filling in the EuroQol questionnaire.* In: Sintonen H (ed). EuroQol Conference Proceedings 1992. Kuopio (Finland): Kuopio University Publications E. Social Sciences 8, 1993.

59 Loomes G. *The myth of the HYE.* Journal of Health Economics 1995;14 (in press).

60 Bleichrodt H. *QALYs and HYEs: under what conditions are they equivalent?* Journal of Health Economics 1995;14 (in press).

61 Michel BC, Bonsel GJ, Stouthard MEA, McDonnell J, Habbema JDF. *Long term cost-effectiveness analysis of liver transplantation; The Groningen liver transplantation programme 1979-1991.* (in Dutch). Ned Tijdschr Geneeskd 1993;137:963-969.

62 Hout BA van, McDonnell J. *Estimating a parametric relationship between health description and health valuation using the EuroQol instrument.* In: Björk S, ed. EuroQol Conference Proceedings. Lund (Sweden): Institute for Health Economics, 1992.

9

Valuation of health states by the general public: feasibility of a standardized measurement procedure (the Bergen op Zoom survey)

Originally published as:
Essink-Bot ML, Bonsel GJ, Van der Maas PJ. Valuation of health states by the general public: feasibility of a standardized measurement procedure. Soc Sci Med 1990;31(11):1201-1206. Reprinted with permission from Elsevier Science Ltd.

9.1 Abstract

In the context of an international collaborative study (the EuroQol enterprise) we tested the feasibility of a procedure to measure valuations of health states in the Dutch general population. A postal questionnaire was sent to a random sample of 200 households in a town in the Netherlands (± 50,000 inh.). Respondents were requested to value 14 six-dimensional health states by means of visual analogue scaling (VAS). The response was considered as satisfactory (57%) given the demanding task and the response-rates to postal questionnaires generally observed in the Netherlands. However, about a fifth of those willing to complete the questionnaire did not manage to use a VAS to express their opinion. Inconsistent answers were relatively rare. Generally consensus existed with regard to relative (ranking) and absolute values of different health states. These first results have encouraged us continue with the development of this international instrument for the valuation of health states.

9.2 Introduction

Cost-effectiveness analysis and cost-utility analysis have currently been recognized as sources of information for decisions about the incorporation of new medical technologies in health insurance schemes.[1,2] For example, governmental decisions about the reimbursement of heart transplantation in the United States, the United Kingdom and the Netherlands were taken using the results of national studies on the costs and effects of this intervention.[3,4,5] Despite a growing interest in the measurement of costs and effects of health care intervention programmes, published studies continue to display large disparities with respect to concepts and operational design, especially in the measurement of effectiveness.[1,6]

In cost-utility analysis, effects on length of life and on health status are represented in a composite measure, e.g., quality adjusted life years (QALYs). One of the strategies to obtain values that can be used to combine life years and quality is empirical measurement of valuations of health states.[7] The operational designs of various empirical techniques to determine the quality adjustment index show many differences. This is to some extent undesirable. The application of different descriptive systems and different operationalizations of valuation methods limit the possibility to benefit from methodological studies conducted by others elsewhere, especially in the judgement of aspects of validity.[1] This hampers the progress and development of the resarch field. Furthermore, it leads to incomparability of results of cost-utility analyses of different intervention programmes, thus precluding the intended use of this information in setting priorities among interventions by governments, health insurance companies and others.

Following an initiative by Professor Alan Williams (York) in 1987, several European research groups combined their research efforts to form the EuroQol Group, with two principal aims: firstly, the development of a 'common core' of methods and practical devices to measure health state valuations, and secondly, the establishment of a common set of data collected with this 'common core' instrument in different European countries. Details on the aims and the development of this research group were published elsewhere.[8] Based on the shared experience and on the results of pilot studies, consensus was reached about a prototype of a common instrument to measure valuations on health states in 1988. An international pilot study using this measuring instrument has been conducted since.

This article describes the results of the Dutch contribution of the international pilot study and addresses the following questions:

1. Is the proposed procedure for measuring valuations on health states feasible for large scale surveys? Feasibility in this context should include the following aspects:
   - The feasibility of valuing complex multi-dimensional health-state descriptions.
   - The feasibility of the valuation of health states by the general public, as opposed to students or well-educated convenience samples.[9,10]
   - The feasibility of the valuation of health states by means of a postal questionnaire, as opposed to the more commonly used interviewer-supported designs.[10]
2. What are the actual values of health states, that range from the health state of, for example, heart transplantation candidates to that of the healthy population?

3 Is there sufficient consensus among respondents with respect to valuations of health states to justify future research to be directed at application of the results in cost-utility analysis?

Questions relating to the influence of background variables (including nationality) on the valuations of health states, validity of the measurement procedure and use of the results in cost-utility analysis will be addressed more extensively in later papers, combining the results of several national studies.

9.3 Methods

9.3.1 *Evaluative health status measurement*

A procedure to empirical evaluative health status measurement consists of three consecutive steps:

1. descriptive health status measurement of the target population, usually patients;
2. valuation of the resulting health state descriptions by subjects who represent, for

example, the general public, experts or patients;
3. combination of the health state valuations with survival data, for example into QALYs.

9.3.2 *The concept of health used*

In the operationalization of health status, a multidimensional approach was adopted in order to take the complexity of the concept of health into account. A 6-dimensional (6D) concept of health status was agreed. This is opposite to, for example, the approach of Rosser and Kind who used a two-dimensional operationalization of health status, i.e., disability and distress.[11] The choice of dimensions was guided by a careful review of existing descriptive health status measures such as the Nottingham Health Profile and the Sickness Impact Profile, and of the operationalizations of health status used by Patrick and Bush and by Rosser & Kind respectively.[9,10,12,13] Each dimension was divided into levels or categories. Each level represents a different degree of difficulty with respect to that specific dimension. Three dimensions were divided into three categories, the other three into two categories, see Table 9.3.2.1. Thus a complete health-state description or 'vignette' consists of six statements ('items'). Theoretically this set of dimensions and items allows for 216 (2^3 x 3^3) permutations. Each possible vignette can be characterized by a string representing the item-levels per dimension, '1' representing the optimal category, '2' and '3' representing the intermediate and worst categories, respectively, in the case of a dimension divided in 3 categories. The Figure '2' represents the worst category in the case of a dimension divided in 2 categories. An example of vignette is shown in Table 9.3.2.2.

9.3.3 *The EuroQol valuation questionnaire*

A questionnaire was developed for the valuation of health-state descriptions (step 2 of the procedure mentioned above) by members of the general public.

The EuroQol Group selected a standard set of thirteen health state descriptions to be valued in the questionnaire. The state of 'being dead' was added.[14] Two states were presented twice. This resulted in sixteen vignettes.

The vignettes were presented in boxes on 2 pages of the questionnaire. Four boxes were placed on either side of a vertically placed visual analogue scale (VAS) in a random sequence. The endpoints of the VAS were marked with the words 'worst imaginable health state' (0) and 'best imaginable health state' (100). No additional information about the interpretation of the scale was added. Respondents were asked to indicate how good or how bad each state was to them by drawing a line from each box to the thermometer. The duration of each state was stated to be 1 year; what happens afterwards was stated not to be known. A detailed instruction paragraph was added, including an illustration of the valuation method using a non health-related example. The lay-out of the questionnaire was carefully designed.

TABLE 9.3.2.1 Dimensions of the EuroQol health concept (6D)

Mobility (3 levels)
Daily activities and self care (3 levels)
Working performance (2 levels)
Family and leisure performance (2 levels)
Pain/discomfort (3 levels)
Mood (2 levels)

TABLE 9.3.2.2 Example of a vignette: EuroQol state 112232

No problems in walking about (level 1)
No problems with self-care (level 1)
Unable to perform main activity (level 2)
Unable to pursue family/leisure activities (level 2)
Extreme pain or discomfort (level 3)
Anxious or depressed (level 2)

The main task of rating the health-state descriptions in the boxes on the VAS was preceded by the task of classifying and rating the 'own health state' of the respondent. Questions about bakground characteristics that might influence the rating of health states (including age, educational level, experience with illness, own state of health) were presented at the end of the questionnaire.

9.3.4 *The sample*

The questionnaire was mailed to a random sample (n=200) of the general population in December 1988, followed by a reminder two weeks later. Sampling was based on postal codes.

9.4 Results

Responses were obtained from 112 persons. The response rate (excluding 4 deceased persons) was 57%, assuming all addresses were correct, which is probably optimistic. Background data on the respondents are shown in Table 9.4.1. Comparison with data relating to the population that was sampled for the study showed some differences in age- and sex distribution. The relative overrepresentation of men among the respondents may be due to the method of sampling, as the questionnaire was directed at the administrative head of each household.

Five respondents returned their questionnaire blank. Twenty-one subjects (20%) of the remaining 107 clearly had not understood the task, i.e., the use of a VAS to express their opinion. This response was chiefly found among the older and less educated respondents (Chi-square: age (2 strata) p=0.003, education p=0.001). These respondents were left out of further analysis. Five of the remaining 86 completed only one of the two pages on which the valuation task was presented, overtly because they thought the second page to be a replication of the first. Data from these five were included.

TABLE 9.4.1 Background data of respondents (n=86); some background data of the population (age > = 15)

Variable	Respondents	Sample population
Age		
15-29	16 (19%)	30%
30-44	32 (37%)	28%
45-59	11 (13%)	19%
60-70	42 (26%)	15%
> = 75	5 (6%)	7%
Sex		
Male	54 (63%)	52%
Education		
Minimum	37 (43%)	55%
Intermediate	34 (40%)	31%
Higher/degree level	15 (17%)	14%
Main activities		
Employed	46 (54%)	
Retired	19 (22%)	
Housework	15 (17%)	
Student	1 (1%)	
Incapacitated	4 (5%)	
Seeking work	1 (1%)	
Rating of own health		
< 80	18 (21%)	
80-90	29 (34%)	
> 90	32 (37%)	
Missing	7 (8%)	

To assess aspects of the feasibility of the questionnaire, respondents were questioned about the difficulties they experienced in answering it. Forty-three percent of the respondents judged the questionnaire as being very (6%) or rather (37%) difficult, while 57% reported it to be fairly (45%) or very (12%) easy. The respondents needed a mean time of 20.3 minutes (SD 12.4 minutes) to complete the questionnaire.

Respondents rated their own health status on the VAS (range: 0-100) with a mode of 85, a median of 85 (interquartile range 8) and a mean of 81 (SD 18). Indeed, those who classified themselves on all predefined dimensions as being in the best category (11111) (n=52), attributed a significantly (p < 0.01) higher value to their own state of health (mean 89, SD 7) than those who reported a suboptimal level on any dimension (n=40; mean 70, SD 22).

The results of the valuation of the 16 selected health states are summarized in Table 9.4.2. The sequence of the states follows the median scores, as it may be assumed that the measurement level of the valuations is at least ordinal. Ranking following the arithmetical means does not result in any change in the sequence, however.

The dispersion of the attributed values was large, especially for 'bad' health states. 'Being dead' yielded a heterogenous response, the range of attributed values ranging from 0 to 100. Seven of 80 respondents (9%) valued 'being dead' equal to or higher than 50. Two health states were presentend twice (112222 and 'being dead'). Scores were compared for both pairs (Spearman's rank correlation coefficient 0.76 and 0.95 respectively; Pearson's correlation coefficient 0.69 and 0.94 respectively). The differences between the mean ratings were statistically insignificant ($p > 0.05$ for both pairs) (see Table 9.4.2).

TABLE 9.4.2 Valuations for 14 health states

Health state[1]	Mode	Med.	I.Q.[2]	Mean	S.D.	n
111111	100	95	5	92	14	86
111121	80	85	10	81	19	82
111112	85	78	10	73	21	81
111122	70	70	13	69	21	86
112121	60	65	15	64	22	85
112131	65	60	14	55	23	83
112222(a)[3]	30	43	13	42	21	86
112222(b)	40	40	13	41	21	82
112232[4]	25	33	11	37	22	81
212232	20	20	8	26	20	86
being dead (a)[3]	0	5	23	21	26	80
being dead (b)	0	5	20	19	25	77
222232	5	8	6	12	15	85
232232	0	6	4	11	16	83
322232	0	5	5	10	17	81
332232	0	4	5	7	12	85

[1] for clarification of representation of health states by strings of numbers see section 9.3.2

[2] Med. = median score, I.Q. = Interquartile Range.

[3] These two states ('112222' and 'being dead') were presented twice in the questionnaire.

[4] The string '112232' represents the vignette presented in Table 9.3.2.2.

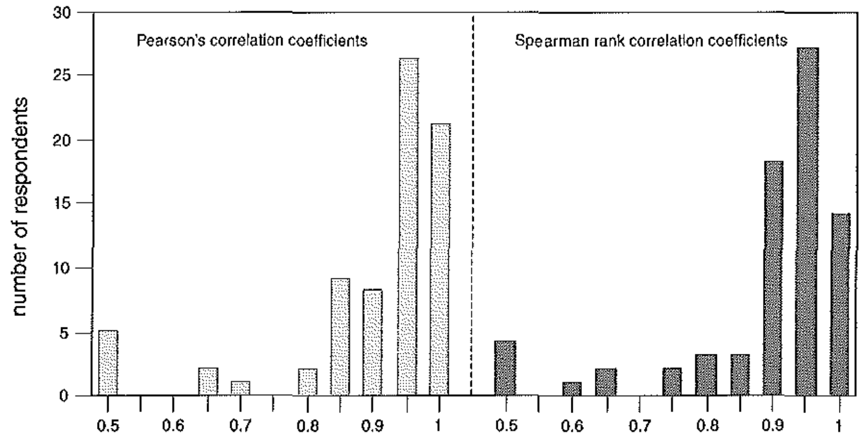
We then examined the logical consistency of the ratings, both at a group level and at an individual level. Consistency of valuations may be checked by comparing values attached to pairs of health-state descriptions with a logical order; i.e., the one health-state description being dominant over the other. For example, 112222 is expected to be assigned a higher (or at least equal) rating than 212232.
In the case of non-dominant pairs, e.g., 112131 and 112222, no logical ranking order exists. 'Being dead' was excluded from the analysis of the consistency of the ratings. At a group level, the median scores were completely consistent (see Table 9.4.2). However, at the individual level inconsistencies occurred. The 14 health states account for 82 dominant pairs. Overall, only 5% of the answers proved to be inconsistent at individual level. Illogical ranking occurred more often as pairs of health states were more alike, see Table 9.4.3.

The degree of consensus among the respondents with respect to the ratings was examined, firstly by comparing individual rankings with the group ranking, as the measurement level was assumed to be at least ordinal. As cardinal measuring properties are conceivable, a second comparison of individual ratings with the group means was also carried out.The results are shown in Figure 9.4.

TABLE 9.4.3 Consistency of the ratings, by comparing ranks assigned to 82 dominant pairs of health states

Distance between 2 health states	Number of pairs	Inconsistent ratings
1	14	16.3%
2	14	5.9%
3	13	2.8%
4	12	1.8%
5	9	1.7%
6	8	1.7%
7	7	1.4%
8	4	1.2%
9	1	1.2%
Total	82	5.0%

FIGURE 9.4 Correlation of individual scales with group scale (n=74)



9.5 Discussion

We examined the feasibility of a measurement instrument to elicit valuations for a number of complex health-state descriptions from the general public. The design of the instrument was based on previous experience and supporting pilot studies by the EuroQol Group, a collaborative group of European researchers.[15,16,17] Essentially the task consisted of the valuation of six-dimensional health-state descriptions on a vertical rating scale in an unsupported situation (postal questionnaire). Arguments in favour of the feasibility of the procedure are:

- the relatively high response, taking into account the demanding nature of the questionnaire;
- the acceptable level of experienced difficulty;
- the logical ranking of the health states by the respondents as a group.

We concluded that rating of health states on a rating scale by postal questionnaire might be feasible. However, inappropriate response did occur (20%), and was related to age and level of education. Improvement of the instruction paragraph will probably enhance completion rates. A non-response study will be undertaken to investigate whether subjects who did not respond at all (non-respondents) and subjects who were willing to respond but do not succeed (unsuccessful respondents), respectively, differ from successful respondents with respect to the variables of interest, i.e., valuations of health states.

Judging from the examination of the logical consistency of the ratings, the valuation task was, in general, reasonably well understood. The finding that respondents who classified their own health better also rated it significantly higher is another indication that the subjects interpreted the valuation task correctly.

The dispersion of the assigned values was fairly large. However, psychophysical methods typically yield higly variable observations.[18]

The consensus found among respondents with respect to the ranking of the health-states is important.[19] It supports earlier evidence of homogeneity of society with regard to valuations of health states.[11] However, the possibility of bias introduced by the response rate should be kept in mind. A comparison of the results of this pilot survey with preliminary results of similar studies in other European countries showed virtually the same ranking order of health states.[8]

Though these provisional results seem to justify further experimentation with the EuroQol valuation questionnaire, there remain many questions to be answered. The reliability and aspects of the validity of the questionnaire should be determined. Validity testing should include comparison with other valuation methods (time trade-off, standard gamble). Furthermore, the generalizability of the results to subgroups of the population, e.g., patients, health care workers, should be investigated. For application of the health-state valuations in cost-utility analysis at least an interval scale is required.[6,20] Therefore, the measurement level of the VAS values should be established. Are respondents merely ranking the states, or are the numbers on the scale interpreted quantitively? If the health states appear to be only ranked by the respondents, a scaling procedure could be performed in order to achieve cardinal values. Even if the measurement level is interval, the values resulting from the measurement procedure should probably not be used directly as a quality index in the computation of QALYs. The relation between the elicited valuations for health state as measured by time trade off or standard gamble may be e.g., logarithmic or linear.[6,18,21] Probably a transformation procedure should be employed before using the results in cost-utility analysis.[22] More generally, the way to apply the results in QALY calculations should be determined.

Problems remain with the valuation of 'being dead'. The presentation of 'being dead' as a health state seems to cause opposition in the respondents. This might be the result of conflicting interpretations of 'being dead'; for example 'absence of life' or 'the process of dying', or even an interpretation as a type of reference point ('something very bad'). This multi interpretability might be the cause of cognitive problems when comparing 'being dead' with health states during the completion of the questionnaire.

Though many questions have yet to be answered, the first results with the EuroQoL-instrument are sufficiently encouraging to continue its development as a standardized measure of obtaining valuations of health states.

Note

1. The 6D, 2/3 level EuroQol was changed into the present 5D, 3L after the first pilotstudies.

References

1 Drummond MF, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes.* Oxford University Press, Oxford, 1986.

2 Leaf A. *Cost effectiveness as a criterion for Medicare coverage.* N Eng J Med 1989;321(13):898-1989.

3 Evans RW, Manninen DL, Overcast TD, Garrison LP, Yagi J, Merrikin K, Jonsen AR, et al. *The National Heart Transplantation Study: Final Report.* Seattle, Washington: Battelle Human Affairs Research Centers, 1984.

4 Buxton M, Acheson R, Caine N, Gibson S, O'Brien B. *Costs and benefits of the heart transplant programmes at Harefield and Papworth hospitals: Final Report.* Department of Economics, Brunel University and Department of Community Medicine, University of Cambridge, 1985. ISBN 0 11 321033 7.

5 Charro FT, Bonsel GJ, Hout BA van. *Costs and effects of hearttransplantation.* Volume 9: final report (in Dutch). Department of Economics, Faculty of Law, Erasmus University Rotterdam, The Netherlands, 1988. ISBN 90 72245 19 9.

6 Torrance GW. *Measurement of health utilities for economic appraisal.* A review. J Health Econ 1986;5:1-30.

7 Lipscomb J. *Value preferences for health: meaning, measurement and use in program evaluation.* In : Kane RL, Kane RA (eds.). Values and long-term care. Toronto: DC Heath and Company, 1982. ISBN 0 666 04685 x.

8 EuroQol Group. *EuroQol - a new facility for the measurement of health-related quality of life.* Health Policy 1990;16:199-208.

9 Patrick DL, Bush JW, Chen MM. *Toward an operational definition of health.* J Health Soc Behav 1973;14:6-22.

10 Kind P, Rosser R. *The quantification of health.* Eur J Soc Psychol 1988;18:63-77.

11 Rosser R, Kind P. *A scale of valuations of states of illness: is there a social consensus?* Int J Epidemiol 1978;7:347-358.

12 Hunt SM, McEwan J, McKenna S. *Measuring health status.* Croom Helm, London, 1986. ISBN 0 7099 3584 6.

13 Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. *The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure.* Int J Health Serv 1976;6(3):393-415.

14 Sintonen H. *An approach to measuring and valuing health states.* Sco Sci Med 1981;15(2):55-65.

15 Williams A. *Economics of coronary bypass grafting.* Br Med J 1985;291-326.

16 Buxton M, Ashby J, O'Hanlon M. *Valuation of health states using the time trade-off approach: report of a pilot study relating to health states one year after treatment for breast cancer.* Health Economics Research Group, Brunel University, Discussion Paper no. 2, London, 1986.

17 Bonsel GJ, Bot ML, Boterblom A, Veer F van 't. *Costs and effects of heart transplantation.* Volume 2C: Quality of life before and after heart transplantation - Results (in Dutch). Institute of Public Health and Social Medicine, Erasmus University Rotterdam, The Netherlands, 1988.

18 Stevens SS. *Psychophysics.* John Wiley, New York, 1975.ISBN 0 471 82437 2.

19 Hilden J. *The nonexistence of interpersonal utility scales. A missing link in medical decision theory?* Med Decis Making, 1985;5(2):215-228.

20 Weinstein MC. *Economic assessments of medical practices and technologies (tutorial).* Med Decis Making 1981;1(4):309-330.

21 Llewellyn-Thomas HA, Sutherland J, Tibshirani R, Ciampi A, Till JE, Boyd NF. *Describing health states. Methodologic issues in obtaining values for health states.* Med Care 1984;543-522.

22 Torrance GW. *Multiattribute Utility Theory as a method of measuring social preferences for health states in long-term care.* In: Kane RL and Kane RA (eds.). Values and long-term care. D.C. Heath & Company, Lexington (USA), 1982. ISBN 0 666 04685 x.

10

Generalizability of valuations on health states collected with the EuroQol questionnaire (the Rotterdam survey)

Published as:
Essink-Bot ML, Stouthard MEA, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQol questionnaire. Health Economics 1993;2:237-246. Reprinted with permission of the publisher.

10.1 Abstract

Problems. Non-response and non-usable response were found in population surveys on valuation of health states. If non-response is selective regarding valuations, then generalization of the resulting values to the whole survey population is not permitted. This could limit the use of empirical utility values in resource allocation in health care.

Methods. Response behaviour of a sample of 1400 from the Dutch general population to the mailed EuroQol-questionnaire was analyzed by four methods. I. Phoning resolute non-respondents; II. comparison of zip code characteristics of respondents and non-respondents (because individual data on background characteristics were not available for the non-respondents); III. analysis of response over time (wave-analysis); IV: comparison of background variables of successful (less than two valuations missing) and unsuccessful respondents, combined with analysis of the effect of these background variables on valuations.

Results. No indications for selective non-response were found, although the phenomenon appeared hard to investigate. The successful response came from a slightly younger and better educated subsample. However, a general influence of age and educational level on valuations could not be shown. This finding is consistent with the literature.

Conclusion. Although the existence of selective non-response cannot be excluded, its relevance can be considered small. This finding is encouraging for the use of empirical utility values in allocative decisions.

10.2 Introduction

Health policy makers facing explicit allocative decisions have recognized economic evaluation as a possible source of information. Ideally, economic evaluation enables health policy makers to rank health care services according to their relative efficiency. This information can be helpful in organizing the thoughts in the process of priority setting, although for definite choices additional information, e.g., about the distribution of costs and effects, is equally necessary.[1]

Cost-utility analysis (CUA) is generally the preferable form of economic analysis, because it takes into account the value of health outcomes in alternative programmes.[1] The empirical measurement of utilities is not straightforward. An approach in two stages is common. In the first stage, health-state descriptions are obtained from patients. In the second stage, these health-state descriptions are valued. The most important methodological choices to be made are the descriptive system for health status, the valuation method and the subjects who will perform the valuation task. The descriptive system should be non-disease specific in order to enable comparisons across programmes and across diagnoses. As for valuation methods, the feasibility of classical methods like standard gamble and time trade-off in large population surveys is questionable.[2] As for as whose valuations should be used, we think that in the case of public policy decisions the societal viewpoint should be taken, so the values of the general public are those we should use.[1,3]

Since 1987, the EuroQol Group has been developing an internationally standardized, feasible, valid and reliable method for the measurement of the general public's valuations on non-disease specific health outcomes. A postal questionnaire for the measurement of valuations on different health states has been developed.[4] The results of pilot studies with the EuroQol instrument in the UK, Sweden and the Netherlands had several features in common.[5,6,7] First of all, the questionnaire appeared to be practically feasible. Secondly, the resulting valuations were remarkably similar; the international interchangeability of valuations had not been seriously investigated before. Another shared result was the fairly high percentage of non-response and unsuccessful response, raising the question of generalizability of the values. For the Dutch pilot survey, the non-response rate was 43%; however, 20% of the response turned out to be unusable, despite careful questionnaire design.[5] If the valuations on health outcomes of respondents differ from those of non-respondents (selective non-response), serious problems arise concerning the generalization of the valuations.

Non-response is found in any population survey. From an economic perspective the phenomenon is of special interest. If non-response is selective regarding the relevant variables (valuations), should we make significant efforts to collect valuations on health states from non-respondents? Or should their views be disregarded, as they do not use the opportunity to have their say in this matter? This issue precedes the issue of aggregation of individual preferences.[8]

In conclusion, there appeared to be enough reasons to undertake a thorough investigation of the response behaviour to the EuroQol questionnaire, which is the principal issue of this paper. A few words will be devoted to consensus among respondents.

10.3 Material & Methods

10.3.1 The study design

The study population consisted of a random selection of 1400 households in Rotterdam, the Netherlands. Respondents were approached by a mailed questionnaire in January 1991. Reminders were sent two weeks (card) and three weeks (whole questionnaire) later. The actual respondent in each household was randomly selected by addressing the accompanying letter to the first adult (≥ 18 years) member of the household who would next celebrate his or her birthday.

Four approaches were used to analyze response behaviour.

1. The only essentially valid method to judge selectivity of response is comparison of respondents and non-respondents *on the variables of interest*, i.e., valuations on health states. We therefore tried to obtain these from the people in the sample who did not respond on the mailed questionnaire. A crude response rate of at least 90% was pursued in a random 350 subsample by means of phoning all resolute non-respondents, if necessary repeatedly, 5 weeks after the first mailing of the questionnaire. People answering the phone were asked to complete the questionnaire. If they refused, we asked them why, and tried to obtain data on background variables.
2. A secondary approach to the analysis of non-response is comparison of respondents and non-respondents on background variables. Our study population was a general population sample, so we had no source of data on background variables on the individual level except the questionnaire, which was not answered by non-respondents (because if they did, they were respondents). However, there was external information available on both respondents and non-respondents on the level of zip-code areas. It is stressed here that these zip code characteristics are aggregated data: a zip code area in the Netherlands consists on average of 15 households. Examples of the zip-code characteristics we could dispose of are: average purchasing power index, age distribution, household composition etc. Because comparison of respondents and non-respondents on background variables was only possible on aggregate data, the result of this analysis should be interpreted with caution. Recently, the validity of the use of zip code characteristics as proxies for individual data has been shown to be satisfactory (CTM Schrijvers, Department of Public Health, Erasmus University Rotterdam; personal communication). The method using zip code characteristics to analyze non-response has been applied before by De Leeuw.[9]
3. The third approach was analysis of response over time by wave analysis. If the valuations of early and late respondents were different, due to, e.g., cognitive difficulties, and if the reasons for late response were partially the same as for non-response, then the valuations of respondents and non-respondents could be assumed to be different. Three groups of respondents were identified by our mailing actions. Early respondents (questionnaire received within 3 weeks after the first mailing) were assumed to have responded to the first mailing, medium respondents (3-5 weeks) to the first reminder card and late respondents (>5 weeks) to the second mailed questionnaire. Differences in valuations between these groups were analyzed by one-way analysis of variance. Because the data were not normally distributed and because the nature of the data is probably quasi-interval we used Kruskal-Wallis analysis of variance by ranks.
4. The finding that not all people who were willing to complete the questionnaire in fact succeeded in doing so was recognized as a separate problem. To analyze selectivity of successful response, background characteristics as reported in the questionnaire by successful and unsuccessful respondents were compared. Successful (= usable) response was defined as only one or two valuations

missing, the assumption being that if only one or two states were missing the respondent had essentially understood the task. If non-usable response was selective regarding background characteristics, and if these background characteristics were of influence on valuations on health states, this would be strongly indicative of selective non-response regarding valuations. The effects of a set of background characteristics (i.e., sex, age, educational level and global evaluation of own health state; this last item being operationalized on a vertical visual analogue scale with marked endpoints, '0' labelled as 'worst imaginable health state' and '100' labelled as 'best imaginable health state') of respondents on the valuations were studied by Kruskal-Wallis one-way analysis of variance by ranks because of the qualities of the data mentioned above. A limitation of this non-parametric technique is that it does not allow for the simultaneous estimation of the effect of an independent variable (each background variable) on a set of dependent variables (16 health state valuations). Therefore, a multivariate analysis of variance (MANOVA) on the valuations was also carried out (one-between one-within repeated measurements design; Wilks' lambda).[10,11] The MANOVA was performed using the sequential approach for non-experimental designs in the SPSS computer package.

10.3.2 *The EuroQol concept of health status*

The revised EuroQol concept of health status is shown in Table 10.3.2. With this concept, a health-state description can be composed by taking one level for each dimension. For example, state 11231 indicates a state of health without mobility or self-care problems, some problems with usual activities, extreme pain or discomfort, but no anxiety or depression.

The EuroQol concept of health status theoretically allows for 3^5 or 243 composite health-state descriptions.

10.3.3 *The EuroQol questionnaire*

Respondents are asked to classify their own state of health using the EuroQol-concept on the first page.

They are then asked to rate their own overall state of health on a 'thermometer', i.e., a vertical visual analogue scale (VAS) with marked endpoints: 0 = 'worst imaginable health state' and 100 = 'best imaginable health state'.[12]

The core task of the questionnaire is the valuation of 16 composite health-state descriptions, concerning 'someone like you', on a VASC as described above. The duration of the health states is stated to be one year; what will happen afterwards is stated not to be known. The 16 health-state descriptions are presented on two pages A and B. One page of the valuation task is shown in Figure 10.3.3.

Data on background characteristics (e.g., age, sex, education, experience with illness) were collected on the last two pages of the questionnaire.

The standard EuroQol questionnaire contains a fixed selection of 14 different health states. Health states '11111' and '33333' are presented on both valuation pages of the questionnaire. In the present study, 14 additional health states were selected to be valued. Two new valuation pages (C, D) were created; the additional health states were assigned randomly to each of them. Four versions of the questionnaire were constructed, namely AB (standard EuroQol), CB, AD and CD.

All health states occurred in two versions of the questionnaire, except 11111 and 33333, which occurred twice in each version. The four versions of the questionnaire were distributed randomly among the addressees in the sample.

TABLE 10.3.2 The EuroQol concept of health (5D - 3L)

Mobility	
1	No problems in walking about
2	Some problems in walking about
3	Confined to bed
Self Care	
1	No problems with self care
2	Some problems washing or dressing
3	Unable to wash or dress self
Usual Activities (e.g., work, study, housework, family or leisure activities)	
1	No problems with performing usual activities
2	Some problems with performing usual activities
3	Unable to perform usual activities
Pain/discomfort	
1	No pain or discomfort
2	Moderate pain or discomfort
3	Extreme pain or discomfort
Anxiety/depression	
1	Not anxious or depressed
2	Moderately anxious or depressed
3	Extremely anxious or depressed

10.3.4 Presentation of the results

The measurement level of a VAS is assumed to be quasi-interval. Therefore, the resulting valuations are presented by the median as well as by the mean. On the level of the respondents as a group, consensus is determined by examining the frequency distributions of the Spearman rank correlations between individual rankings and the group ranking. Only complete data could be used for this analysis.

10.4 Results

10.4.1 Response

Five questionnaires were returned in the original envelope as 'undeliverable'. A total of 980 questionnaires were returned in the prestamped return envelope, yielding a 70% (980/1395) crude response rate. As 111 questionnaires out of 980 were returned blank, the non-blank response-rate was 62%. Non-blank response rate per version amounted 66%, 55%, 64% and 62% for version AB, CB, AD and CD respectively. In the following, 'response' means 'non-blank response'.

Returned blank questionnaires were considered to be indicative of unwillingness to participate for some reason.

10.4.2 *Valuations; consensus*

The resulting valuations on health states are presented in Table 10.4.2. Data from successful respondents as defined above were used. The states are ordered according to the medians. The data are presented pooled as no differences in valuations between versions of the questionnaire could be proved.

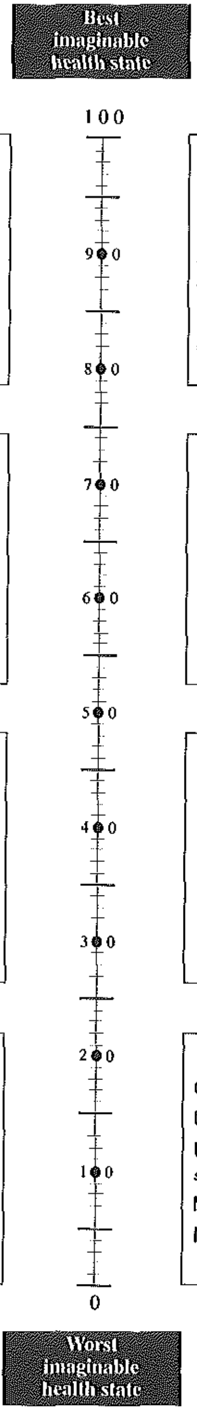
Consensus refers to the extent to which respondents agree on the valuation of health states. The frequency distribution of correlations between individual rankings and the group ranking is shown in Figure 10.4.2. The percentage of respondents with a rank correlation with the group ranking lower than 0.50 was 4.5%.

TABLE 10.4.2 Valuations of health states for respondents with two or less missings (n= 643)

Health state	Median	Mean	S.D.	*n*
11111b*	97	92.3	13.2	639
11111a*	97	92.3	13.6	639
11211	80	80.5	14.4	331
11121	75	73.6	18.5	332
11112	75	73.4	18.7	341
12111	70	67.9	23.7	337
21111	68	62.9	23.2	333
11221	65	65.5	18.1	300
11122	60	60.0	20.7	333
21211	58	52.9	23.5	306
12212	54	52.7	20.2	297
21212	50	48.5	20.3	300
32211	45	45.2	23.3	338
21232	30	35.1	23.9	333
23223	30	29.9	22.6	308
22233	20	27.1	23.2	333
33321	20	26.3	23.0	332
22323	20	26.0	23.0	339
32233	20	24.9	23.4	307
22333	20	24.8	22.8	306
23332	15	21.2	21.3	306
32333	15	20.8	22.7	297
33332	15	20.7	21.9	308
33233	15	19.8	21.5	296
23333	10	15.6	20.1	299
33333b*	5	14.4	23.2	642
33333a*	5	13.3	23.1	642
unconscious	4	16.1	27.3	336

* health states presented twice

FIGURE 10.3.3 Page 1 of the valuation part of the questionnaire



No problems in walking about
No problems with self-care
Some problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
No pain or discomfort
Not anxious or depressed

No problems in walking about
No problems with self-care
No problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
Moderate pain or discomfort
Not anxious or depressed

No problems in walking about
No problems with self-care
No problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
No pain or discomfort
Not anxious or depressed

Some problems in walking about
Some problems with washing or dressing self
Some problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
Extreme pain or discomfort
Extremely anxious or depressed

Some problems in walking about
No problems with self-care
Some problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
Extreme pain or discomfort
Moderately anxious or depressed

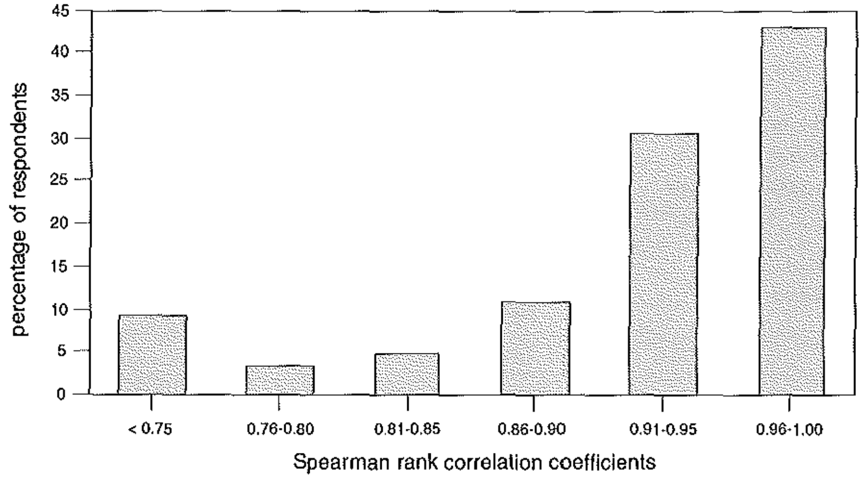
Confined to bed
Unable to wash or dress self
Unable to perform usual activities *(e.g. work, study, housework, family or leisure activities)*
Extreme pain or discomfort
Extremely anxious or depressed

No problems in walking about
No problems with self-care
No problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
Moderate pain or discomfort
Moderately anxious or depressed

Confined to bed
Unable to wash or dress self
Unable to perform usual activities *(e.g. work, study, housework, family or leisure activities)*
Moderate pain or discomfort
Not anxious or depressed

PLEASE CHECK THAT YOU HAVE DRAWN ONE LINE FROM EACH BOX (THAT IS, 8 LINES IN ALL)

FIGURE 10.4.2 Consensus (frequency distribution of rank correlations between individual ranking and group ranking), n = 493



10.4.3 *Analysis of response behaviour*

1. We tried to collect valuations of 92 addressees of a random 350-subsample who had not responded after two reminders, nor returned a blank questionnaire, by means of repeated reminder phone calls. Eightteen of them did not answer their phone even if tried 5 times (19%) while one addressee appeared to have died. Twenty-one promised to return the questionnaire, of which only 9 were actually received. Of the remaining 52, 47 refused to answer any question about their age, educational level etc. As reasons for not participating they offered 'not seeing the sense', 'not being interested' 'principally never participating in surveys', 'too busy'. Eight non-respondents said they did not understand the questionnaire, but refused help to complete it.

   We concluded that a general unwillingness to cooperate in surveys was the main reason for non-response. The number of 9 extra completed questionnaires was considered too small to analyze separately.
2. No relevant differences could be detected between postal area characteristics of people in the sample who returned a non-blank questionnaire and of those who returned a blank questionnaire or nothing at all.
3. The differences in ranking order of states in three response waves (wave analysis) were insignificant at the 5% level.

The results of the wave analysis for the medians of valuations for 4 states are illustrated in Figure 10.4.3 (the results for the other states were analogous). No differences in response over time were detected.

4. Comparison of successful respondents with unsuccessful respondents showed the modal successful respondent to be between 31 and 45 years of age, with medium level education. The modal unsuccessful respondent was older and less well educated, all other measured variables being the same. Among the successful respondents, no significant effects of background characteristics (age, sex, education, valuation of own health, experience with illness) on the ranking of the states could be detected.

   The results of the MANOVA procedure (in which the data are treated as interval data) are shown in Table 10.4.3. The figures in this table can be interpreted as follows.

   As expected from the nature of the valuation task (stimulus scaling task), all tests of the main effect 'health states' are highly significant. The effect of the background variable 'age' in version AB of the questionnaire is the one that attracts attention. Both the main effect 'age' (indicating a difference in level of valuations between age groups for the whole range of health states) and the interaction effect (age * health states) are highly significant (p=0.009 and 0.008, respectively). When inspecting the data, it can be seen that the older age group generally values the health states somewhat higher than the other groups. The interpretation of the interaction effect age * health states is that some age groups value some health states differently. In version AB, the interaction effect age * health states is mainly the result of equal valuation of one particular state by all age groups. However, the finding of significant effects of 'age' on valuations is not confirmed in the other versions of the questionnaire.

   The discrepancy of version AB with the other versions can probably not be explained by differences in the severity of the health states in the questionnaire versions, as the whole range of health states is covered in all four versions. Although somewhat unsatisfactory, the conclusion at this moment must be that the effect of the background variable 'age' on valuations on health states *in general* is not unequivocal. If the findings for version AB are indicative of a systematic age-effect, this should be confirmed in future research.

   The other four significant p-values in Table 10.4.3 concern interaction effects. These findings impress as patchy, and when inspecting the data it is clear that they do not reflect any systematic effect.

   More detailed results of the MANOVA are available from the authors on request. Our conclusion of the MANOVA is that none of the background variables tested has unequivocal significant effects on valuations in general, with a possible exception for age; however, the relevance of this possible age-effect seems to be small.

TABLE 10.4.3 Influence of age, sex, education and rating of own health on EuroQol valuations (per version of the questionnaire); one-between one-within multivariate repeated measure analysis (p-levels).

	AB[a]	CB	AD	CD
Health States (16)[b]	<0.001	<0.001	<0.001	<0.001
Age (4)[c]	0.009	0.84	0.82	0.65
Health States * Age	0.008	0.93	0.50	0.15
Sex (2)	0.08	0.19	0.17	0.27
Health States * Sex	0.87	0.14	0.75	0.05
Education (3)	0.18	0.78	0.75	0.84
Health States * Education	0.26	0.27	0.72	0.04
Own Health[d] (3)	1.00	0.45	0.27	0.43
Health States * Own Health	0.51	0.003	0.01	0.45

[a] AB = standard Euroqol-questionnnaire; further explanation of versions, see section 10.3.3.

[b] Within parentheses: number of levels.

[c] Classified as: level 1<31, 2=31-45, 3=46-60, 4>60 years.

[d] Evaluation of own health state of the respondent on a thermometer from 0 to 100; level 1=0-79, 2=80-90, 3=91-100.

This finding suggests that the selectivity of successful response regarding age and educational level is generally not reflected in the valuations.

The results from approaches 1 to 4 can be summarized as follows: up to now, no selectivity of response has been proven. Successful response came from a slightly selective sample, but this selectivity probably did not influence the resulting ranking order or the mean valuations relevantly.

10.5 Discussion

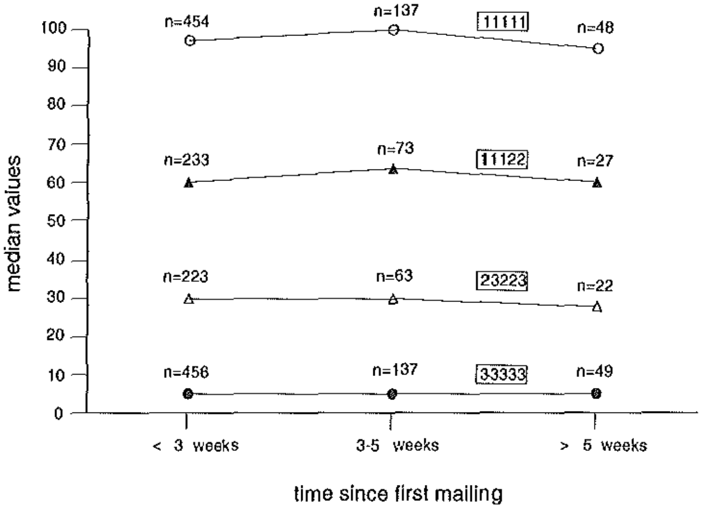
Although a non-blank response rate of 62% is high for a postal questionnaire in a population that is not specially motivated or in any way rewarded, it leaves the investigators without data on the values for health states of the remaining 38%. Therefore, an analysis of non-response and response behaviour was undertaken.

One of the results of this study is the confirmation of the fact that a valid analysis of non-response in a general population sample is hardly a feasible enterprise, just because non-respondents do not respond and external data are not available on the individual level. Essentially, the whole enterprise of testing selectivity of non-response for EuroQol valuations has been unsuccessful.

By application of a set of second-best methods, we found indications that, if non-response is selective, the relevance is probably small.

The percentage of usable response was 46% (643/1400), or 74% (643/869) of non-blank returned questionnaires; the task in the questionnaire appeared to be demanding. In our survey, usable response came from a subsample which was on average younger and better educated. This would pose a threat to generalizability if valuations on health states were influenced by age and educational level, which did not appear to be the case in our data. This is an argument for the acceptability of the selectivity of usable response. We shall try, however, to improve the questionnaire further in order to enable anybody willing to complete the questionnaire to do so.

URE 10.4.3 Response wave analysis (medians) for 4 health states



The finding that the influence of background characteristics on valuations on health states is probably small, is in accordance with the literature.[13] What should have been done if indications for selective non-response relating to valuations had been found? The option of 100% response is unrealistic. If relevant background variables were identified, the opinions of the non-respondents could be estimated by means of a modelling approach. Other approaches are thinkable (e.g., using a random sample of elected politicians), but these are issues beyond the scope of this paper.

As indicated in the introduction, the valuations on health states are meant to be used in CUA. EuroQol valuations are not to be used directly as utility weights; more should be known about the nature of the scores. The measurement level of a VAS is probably quasi-interval. For CUA, at least interval measuring level is necessary. A scaling procedure to establish the meaning of the distances between the numbers should be performed. Furthermore, the exact meaning of the scores should be explored. Do they represent health-state preferences? This question is difficult to answer, as no golden standard for health-state preferences exists, so that criterion validity cannot be evaluated. Construct validity can be investigated by means of multitrait - multimethod analysis.[14,15] Because the health-state descriptions contain a natural ranking order to some extent, each of them could be treated as a separate trait, while the methods should be accepted methods for measurement of health-state preferences. Both the scaling procedure and the construct validity are subjects of current research with the EuroQol instrument.

10.6 Conclusions

Data on health state preferences collected by mail among a sample of the Dutch general population appeared not to be very sensitive to selection bias by non-response. This is an indication that the response-rates as encountered until now are acceptable, and that the results may be generalized to the whole sample, and consequentially, provided the sample was drawn well, to the sampled population.

Acknowledgement

The authors wish to thank Maurice Palmen BSc for his practical assistance, Paul FM Krabbe MSc for performing some of the analyses and for general comment, Professor Paul J van der Maas MD PhD, Ben A van Hout PhD, Erik Nord MSc, and two anonymous referees for their valuable comments on the manuscript, and the members of the EuroQol Group for the inspiring research climate they have created.

References

1 Drummond MF, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes.* Oxford: Oxford University Press, 1987.

2 Froberg DG, Kane RL. *Methodology for measuring health state preferences - I: measurement strategies.* J Clin Epidemiology 1989;42:345-354.

3 Hadorn DC. *The role of public values in setting health care priorities.* Soc Sci Med 1991;32:773-781.

4 EuroQol Group. *EuroQol - a new facility for the measurement of health-related quality of life.* Health Policy 1990;16:199-208.

5 Essink-Bot ML, Bonsel GJ, Maas PJ van der. *Valuation of health states by the general public: feasibility of a standardized measurement procedure.* Soc Sci Med 1990;31:1201-1206.

6 Nord E. *EuroQol: health-related quality of life measurement. Valuations of health states by the general public in Norway.* Health Policy 1991;18:25-36.

7 Brooks RG, Jendteg S, Lindgren B, Persson U, Björk S. *EuroQol: health-related quality of life measurement. Results of the Swedish questionnaire exercise.* Health Policy 1991;18:37-48.

8 Lipscomb J. *Value preferences for health: meaning, measurement, and use in program evaluation.* In: Kane RL and Kane RA, editors. Values and long-term care. Lexington, MA: Lexington Books, 1982;27-83.

9 Leeuw ED de. *Data Quality in Mail, Telephone and Face to Face surveys.* Thesis. Amsterdam: TT-Publications, 1992.

10 Tatsuwoka MM. *Multivariate analysis. Techniques for educational and psychological research.* New York: Wiley, 1971.

11 Tabachnik BG, Fidell LS. *Using multivariate statistics.* New York: Harper Collins Publishers, 1989.

12 Sintonen H. *An approach to measuring and valuing health states.* Soc Sci Med 1981;15(2):55-65.

13 Froberg DG, Kane RL. *Methodology for measuring health state preferences - III: population and context effects.* J Clin Epidemiology 1989;42:585-592.

14 Hadorn DC, Hays RD. *Multitrait-multimethod analysis of health-related quality of life measures.* Med Care 1991;29:829-840.

15 Froberg DG, Kane RL. *Methodology for measuring health state preferences - IV: progress and a research agenda.* J Clin Epidemiology 1989;42:675-685.

11

Conclusions

This chapter presents the conclusions of this thesis in 3 sections. The first section provides a definition of the position of health status as an outcome measure, the second section relates to descriptive health status measurement and the third to evaluative health status measurement.

11.1 The position of health status as an outcome measure

The three cornerstones for the evaluation of the effects of disease and óf medical interventions are survival, health status and disease-specific clinical measures.
Although survival and health status are complementary, their relative importance in outcome measurement is variable. Health status is essential as an outcome measure in the following situations:

- *Effects on survival and health status occurring in opposite directions.* A survival improvement may occur at the cost of an adverse effect on health status. Such a situation occurs for example in the treatment of chronic viral hepatitis with a drug that inhibits virus replication and thus progression of liver damage, but causes severe and lasting fatigue.
- *The absence of a (substantial) survival effect.* Health status improvement is the primary aim of treatment (for example, the treatment of idiopathic urinary incontinence by implantation of a neuromodulation device, if compared with napkin 'treatment').
- *The occurence of different effects within health status.* For example, a drug resulting in improved physical performance but with depression as a side-effect.

Methods for descriptive health status assessment have passed the experimental stage of development. The procedures have matured to such an extent that health status measurement should be a standard part of any research project aiming at the quantification of effects of disease and/or interventions. Ignoring health status in a research proposal, not its inclusion, should be substantiated. This is in accordance with the guideline provided by the Dutch Working Group on Health Status Assessment.
Standardization of research methodology is a prerequisite for use of the results of empirical outcome studies in health policy making. Intended users of empirical outcome results, including health policy makers might play a decisive role in achieving the necessary level of standardization. If outcome data were increasingly used as a basis for decision-making, the imperative need for standardization would become obvious. For example, if the minimal effectiveness of an intervention

required by the 'funnel of Dunning'[1] for inclusion in the basic health insurance package was operationalized as a given improvement in scores from particular health status measures, there would be an obvious impetus to include these measures in medical evaluation studies.

11.2 Descriptive health status measurement

Several instruments for descriptive health status measurement are currently available. The perspective of a study determines what type of instrument is the most appropriate. The health-care policy perspective requires comparability between patient groups and across diagnostic groups. The general importance of this perspective in the evaluation of medical interventions underlies the practical recommendation to employ a combination of measures. A generic measure should be complemented by disease and/or domain specific measures.

For disease- and domain specific health status measurement, use of standard measuring instruments is efficient, for example because the need for norm studies decreases. For generic instruments and QALY-type measures, standardization of the choice of measures is essential. Their 'raison d'être' is the demand for comparability between studies, interventions and patient groups.

The generic instruments currently available are different, firstly, with respect to their operationalization of physical, psychological and social functioning; secondly, to 'testing performance' in different populations of patients; thirdly, to practicality; and fourthly, to specific characteristics (e.g., a link to health status values). From the viewpoint of standardization, implementation of the 'common core' concept (i.e., that all evaluation studies should have at least one measuring instrument in common) may prove to be feasible. A common core measure garantuees a minimal level of comparability.

There is a lack of information about the relative behaviour of the available generic instruments. Parallel research, employing two or more generic measures simultaneously, is one of the methods to provide empirical evidence of the relative value of the instruments.

Sufficiently reliable and validated disease specific instruments to complement generic measures for health status are not easily detected or do not exist for many diseases. In this area a considerable amount of developmental work has to be carried out. A standardized, modular approach, preferably in relation to generic measures, is recommended. Such an approach has been taken, for example, by the developers of the EORTC QLQ-C30, which includes the development of modules specific for certain cancers (e.g., breast cancer, prostatic cancer) and certain treatments (e.g., radiation therapy) to be employed alongside with the 'generic' 30-item core questionnaire.[2] Other examples are available.[3]

The issue of which health status measure(s) should be used generally receives much more attention than other aspects of research design. For example, the importance of the timing of assessments is often underestimated.

Many empirical studies have investigated the health status of relatively 'easy' patient groups: e.g., not too seriously ill, in a chronic, stationary state, not too old, not too young, no cognitive impairments et cetera. Problems in health status assessment occur with respect to, for example, diseases manifesting in attacks (e.g., asthma, migraine); to children; and to psychiatric patients. In fact, attack-type diseases offer a special case with respect to the period of time for which a health status assessment is considered to be representative. The effects of the attacks themselves should be measured as well as their effects on the functioning between attacks, for example by combining 'attack measurement' (by self-assessment immediately afterwards, or assessment by proxy) with measurement of general functioning.

A special characteristic of children compared with adults with respect to health status measurement, is childrens' heterogeneity as a group due to age-related development. This has important consequences, firstly for the *operationalization* of the contents of the physical, psychological and social domain, and secondly for the method of data collection that has to be adapted to the level of communicative and cognitive development. Assessment by (a combination of) proxies and direct observation are the methods available until the child is able to communicate adequately.[4,5] The prognostic value of health status is an issue of special importance in outcome assessment in children.

A similar situation exists with respect to health status assessment among psychiatric patients. Not the concept of health status itself, but its operationalization may be different in somatic and psychiatric patients. The distorted perception of reality, which is part of psychotic disease to some extent precludes collection of valid data on functioning from the patients themselves during psychotic episodes. This situation is to some extent comparable to the relative impossibility to collect empirical health status data from patients themselves during attacks in somatic attack-type diseases. The feasibility, reliability and validity of 'somatic' generic instruments in the psychiatric context remains largely to be investigated.[6] The value of the generally recommended approach regarding the choice of health status measures (i.e., complementing generic measures with disease specific ones to focus on specific aspects of psychiatric patients and psychiatric care) deserves to be investigated.

11.3 Evaluative health status measurement

Comprehensive outcome measurement implies combining health status and survival effects. In the current practice of the disaggregated operationalization of the QALY or QALY type concept [including the disability-adjusted life-year (DALY) used in the 'Global Burden of Disease' project[7]], 'timeless' health status values are combined with life-years. Many authors have questioned the validity of the assumptions underlying this disaggregation. The question at stake is whether the current operationalization violates the assumptions to such an extent that the QALY concept has to be abandoned. Attempts to develop demonstrable superior alternative operationalizations have remained futile for years. Potential users of QALY data should be aware of the consequences of the apparent simplicity of the current operationalization and consequentially of its limitations. They should also realize that cost-effectiveness data expressed as cost per QALY do not yield clear-cut decisions as to how to allocate resources. Such data may be used as one of the sources of information to organize the thoughts in the decision-making processes of policy makers.

The demand for health status values to be applied in the evaluation of medical interventions and in public health models is increasing.[8,9] Because of the lack of a superior alternative, the practical approach of incorporating valuation data on health states in an economic analysis as illustrated in section 8.5 may be defended. This approach was also used in a study evaluating the quality of life effects of breast cancer screening.[10] The application of health status values should be provisionally standardized. The consequent incorporation of the same systematic error in study results is to be preferred above different sources of variability that cannot be disentangled.

This should not detract from the view that evaluative outcome measurement requires further development and refinement, especially with respect to the role of 'time' and to the aggregation of values if different groups in the population are shown to have different value patterns.

Clearly, the empirical work in evaluative health status measurement has not yet reached the same stage of development as descriptive health status measurement. A major multidisciplinary research effort is required, concerning the whole range of conceptual development, empirical testing and the standardization of procedures.

References

1 Committee on Choices in Health Care. *Choose or loose* (Commissie Keuzen in de zorg. Kiezen en delen). Department of Health, 1992.

2 Sprangers MAG, Cull A, Bjordal K, Groenvold M, Aaronson NK. *The European Organization for Research and Treatment of Cancer approach to quality of life assessment: guidelines for developing questionnaire modules.* Qual Life Res 1993;2:287-295.

3 Cella DF, Bonomi AE. *Functional Asssessment of Cancer Therapy (FACT) Scales.* Manual. Chicago: Rush Cancer Institute, 1994.

4 Verhulst FC, Koot M. *Child psychiatric epidemiology.* New York: Sage Publications, 1992.

5 Gemke RJBJ, Bonsel GJ. *Reliability and validity of a comprehensive health status measure in a heterogenous population of children admitted to intensive care.* Accepted for publication in J Clin Epid, 1995.

6 McHorney CA, Ware JE, Lu JFR, Sherbourne CD. *The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups.* Med Care 1994;32(1):40-66.

7 Murray CJL, Lopez AD (eds). *Global comparative assessments in the health care sector.* Geneva: WHO, 1994.

8 World Bank. *World Development Report 1993: investing in health - world development indicators.* Oxford University Press, 1993.

9 Ruwaard D, Kramers PGN, eds. *Volksgezondheid Toekomst Verkenning.* Den Haag: SDU, 1993.

10 Haes JCJM de, Koning HJ de, Oortmarssen GJ van, Agt HME van, Bruyn AE de, Maas PJ van der. *The impact of a breast cancer screening programme on QALYs.* Int J Cancer 1991;49:538-544.

Summary

Disease can ultimately result in a decrease in life span, a decrease in quality of life or a combination of both. Medical care aims at prevention or cure of disease, and when the disease has passed the curable stage, at palliation of the suffering. The ultimate objectives of medical care can be summarized as 'adding years to life and life to years'. In industrialized countries there has been a decrease in the manifestations of acute, life-threatening diseases. Life expectancy at birth is high. Chronic diseases have become an important public health problem. Consequentially, the improvement of quality of life has become the primary objective of medical interventions.

Mortality used to be an important measure to describe the consequences of disease and the effects of treatment. Mortality will of course continue to play an important role as an outcome measure of disease and treatment. However, mortality, or its complement, survival time, as a single outcome measure is often not so informative.

Examples of situations in which the quality of life is at least equally important as an outcome measure can be found easily. If an intervention primarily aims at prolonging life, for example treatment of cancer, a situation may occur where the gain in life expectancy must be weighted against a temporary or permanent decrease in the quality of life. Another example occurs in situations where interventions emerge that aim at saving life, for example certain organ transplantations, and the treatment of children with (otherwise lethal) congenital anomalies. The technical possibilities to prolong life with such interventions inevitably evoke questions about the quality of that life afterwards.

This thesis addresses quality of life measurement in the evaluation of the effects of disease and of medical care. The notion of 'quality of life' in this context has been limited to 'health-related quality of life', or 'health status', defined as quality of life relating to disease and/or treatment. This implies that determinants of quality of life that are not directly related to health or medical treatment are not considered. Comprehensive domains of health status currently include physical, psychological and social functioning.

Chapter 2 provides a global overview of the scientific field of health status measurement. The relationship between conventional clinical parameters (e.g., blood pressure, blood chemistry, E.C.G., X-rays) and health status measures can be described as complementary, each useful in their own context. 'Conventional' medical techniques can, for example, be used to determine the diagnosis of a disease and, because they provide prognostic information, to support treatment decisions. Patient functioning is the variable of interest in the assessment of the ultimate consequences of disease and the effectiveness of interventions (complementary to life span). This applies when doctors treat individual patients as well as when medical interventions are evaluated at

an aggregate level.

Health status can be measured from different perspectives. We distinguished the following: the individual patient's perspective, where the issue is a choice between treatment alternatives; the perspective of groups of patients with similar disease characteristics, where health status measurement provides insight into the effects that are generally to be expected from such interventions; and a societal perspective, where health status information is used to support the decision-making process in resource allocation, mainly in health care. Medical Technology Assessments (MTA) are conducted from a societal perspective, a classical clinical trial from a patient group perspective. It is recognized that research to evaluate the effectiveness of medical interventions purely from the patient group perspective, 'no matter what it costs', is gradually becoming less important.

The perspective (or stated otherwise, the research question) determines the choice of health status measures. We distinguished three main types of health status measures, i.e. generic, disease-specific and domain-specific instruments. Generic measures allow for comparisons of health status irrespective of diagnosis or intervention. Generic health status measurement is a prerequisite in evaluation research from a societal perspective.

Chapter 2 ends by underlining the importance of standardization of health status measurement. Without standardization research results are incomparable. The consequence may be that the results of health status studies in different disease groups cannot be used to rank these diseases according to the relative burden they cause, or that the results of MTAs of different intervention programmes cannot be used to rank these programmes according to their relative (cost-)effectiveness. Another consequence of incomparability of research results is a suboptimal contribution of individual studies to the scientific 'body of knowledge'. Standardization is essential if health status information is to be used in the preparation of and the decision-making in health policy. It is the aim of a nationwide network of researchers engaged in health status assessment, the Dutch Working Group on Health Status Assessment, to promote the standardization process.

Chapters 3 to 7 relate to descriptive health status assessment.

Chapter 3 compares 6 generic health status measures that are available in Dutch, i.e. the SIP, the NHP, the MOS-20, the SF-36, the COOP/WONCA charts and the EuroQol instrument. The concept of health status was operationalized somewhat differently in these instruments. A comparison of testing properties (reliability, validity) based on the literature was not possible due to the population-specificity of testing properties and because the design and reporting of research on testing properties often appeared to be incomparable. However, none of the available measures is superior to the others in all respects. There is a growing need for empirical comparisons of health status measures.

Chapter 4 presents the results of an empirical comparison of the NHP and the SIP when employed in a cross-sectional description of the health status of a group of renal patients treated with haemodialysis. The NHP was found to be more feasible. The NHP scales showed somewhat higher levels of internal consistency. Common factor analysis showed that NHP and SIP data could be efficiently summarized in two higher-order factors - one reflecting physical health, the other mental health. Physical health is emphasized in the SIP, whereas the NHP emphasizes mental functioning.

Chapter 5 presents a similar study involving a sample of migraine sufferers and a control group, in which we compared the feasibility, internal structure, internal consistency, construct validity and 'known groups' validity of 4 generic health status measures (the NHP, the SF-36, the COOP/WONCA charts and the EuroQol instrument). In general, all 4 instruments exhibited a good performance profile. However, both instruments with a multi-item structure performed better than the COOP/WONCA charts and the EuroQol instrument. Test-retest reliability and responsiveness to change over time were not subjects of the comparison. In this study it also appeared to be possible to efficiently summarize the measures in a physical and a mental factor.
Chapter 6 provides an example of applied descriptive health status measurement. The impact of migraine on health status was investigated employing the NHP, the SF-36, the COOP/WONCA charts and the EuroQol instrument in a controlled cross-sectional design. The health status of migraine sufferers appeared to be significantly impaired in comparison to the control group. The difference could only partly be attributed to a higher prevalence of comorbidity, especially self-reported depression, in the migraine group. Migraine has an independent, moderately deteriorating effect on the daily functioning of individuals, in addition to the presumed effects of the attacks.
Chapter 7 reports on the evaluation of the health status effects of liver transplantation in a longitudinal design. For those who survived the hazardous procedure itself, liver transplantation contributed very positively to their health status. Empirical health status assessment in these sometimes very ill patients appeared to be feasible provided the procedure was extremely user friendly and adequate information was supplied to patients, doctors and nursing staff.

Most of the available health status measures, including the NHP and the SF-36, are descriptive instruments. This implies that scores take the form of a profile of scores across the different dimensions of the instrument. We have to go a step further, i.e. to summarize profile scores, if we wish to aggregate the consequences for health status and survival time into one outcome measure. Such a combined outcome measure is needed, for example, in cost-utility analysis and in public health modelling. Summary scores are currently obtained through a procedure in which health status descriptions are valued.
Chapter 8 provides an overview of the field of empirical valuation of health states. The current three-stage approach is illustrated. Patients' health status descriptions (obtained in step 1) are valued in the second step. In the third step valuations and survival time are combined. The results can be expressed as, for example, Quality Adjusted Life-years (QALYs). Important issues of choice in the procedure are addressed. For example, the choice of the group of subjects who perform the valuation task is determined by the research perspective. If a societal viewpoint is adopted, the valuations should reflect this. This is commonly operationalized by obtaining the valuations from a representative sample of the general population (including patients). The valuation method provides another important issue of choice. Visual analogue scaling is advantageous from a practical point of view. Time trade-off may be preferable from the viewpoint of validity. Simple transformation from values obtained by one method into another is theoretically possible if the ordinal ranking of health states is similar.
The operationalization of the QALY or QALY type concept in the current three-stage approach is essentially a disaggregation of the outcomes of a disease or an intervention.

The outcome space can be represented by a tree, where each branch is characterized by a duration, a sequence of health states and a probability of occurrence. The current disaggregated approach is based (among others) on the following two assumptions that are probably seldom completely valid. Firstly, it assumes that the valuation of a health state is independent of what preceded (history) and what will follow (prognosis). It is also assumed that valuation is independent of the duration of the state. Yet a valuation procedure which takes account of history, prognosis and duration is more likely to reflect reality than one which values health states independently. The apparent simplicity of the present operationalization of the QALY concept is one of its strengths. Moreover, a demonstrable superior alternative still has to be developed. We believe that further development and refinement of the approach is justified, provided researchers and potential users of research findings are aware of the limitations such as those described above.

Chapters 9 and 10 present empirical valuation studies in the general population. Both were part of the research programme of the international EuroQol Group, a European network of researchers aiming at scientific methodological progress in the field of the valuation of health states by means of a standardized empirical approach employing the EuroQol instrument. *Chapter 9* presents a pilot study investigating the feasibility of measuring valuations of health states among the general population in a postal survey. The results were promising. However, the rates of non-response and unsuccessful response (especially from elderly and less educated subjects) were considerable. This prompted us to conduct the non-response survey that is presented in *chapter 10.*

Although the phenomena of selective non-response and unsuccessful response appeared hard to investigate, the relevance of such effects for the use of results in policy decision-making seemed to be small. Relevant effects of background variables (age, educational level) on valuations were not found.

Chapter 11 addresses the most important conclusions of this thesis. Health status is the third cornerstone of medical evaluation research, complementary to survival time and disease-specific clinical measures. Methods for descriptive health status assessment have passed the experimental stage. The procedures have matured to such an extent that health status measurement should be a standard part of any research project aiming at the quantification of effects of disease and/or interventions. Ignoring health status measurement, not its inclusion, should be substantiated. Standardization of research methodology is a prerequisite for the comparability and consequently for the use of the results of empirical outcome studies in health policy making. The Dutch Working Group on Health Status Assessment and (potential) users of health status research results, including the government, should cooperate to reach the required level of standardization.

A different picture arises for evaluative health status measurement. Important questions, for example relating to the roles of 'time' and 'sequence', and to the existence of population subgroups with deviating valuation profiles, still need to be answered empirically. The routine use of empirical valuations of health states cannot be recommended without reservations. A major multidisciplinary research effort is still required, covering the whole range of conceptual development, empirical testing and the standardization of procedures.

Samenvatting

Ziekte kan een verkorting van het leven, een vermindering van de kwaliteit van leven, of een combinatie van beide tot gevolg hebben. De geneeskunde beoogt door middel van interventies ziekte te voorkomen of te genezen, en als dat niet kan de schade te beperken en het lijden te verlichten. De uiteindelijke doelstellingen van de geneeskunde kunnen dan ook worden samengevat als 'het toevoegen van jaren aan het leven en van leven aan de jaren'.
In de westerse wereld zijn de manifestaties van acute, levensbedreigende aandoeningen sterk afgenomen. De levensverwachting bij de geboorte is hoog. Chronische, niet acuut levensbedreigende ziekten vormen een belangrijk volksgezondheidsprobleem. Daarmee is verbetering van de kwaliteit van leven het primaire doel van veel medische interventies geworden. Traditioneel was sterfte een belangrijke maat om de gevolgen van ziekte en het effect van behandeling in kaart te brengen. Natuurlijk blijft sterfte een belangrijke uitkomstmaat van ziekte en zorg. Als enige uitkomstmaat is sterfte, of het complement ervan, overlevingsduur, echter vaak niet meer zo informatief. Toevoeging van gegevens over de kwaliteit van leven maakt het beeld vollediger.
Voorbeelden van situaties waarin kwaliteit van leven als uitkomstmaat belangrijk is zijn er legio. In situaties waar levensverlenging wél het primaire doel is, zoals bij sommige kankerbehandelingen, kan het voorkomen dat vanwege het ingrijpende karakter van interventies de verwachte winst in levensduur moet worden afgewogen tegen een tijdelijke of duurzame vermindering van de kwaliteit van leven. Een andere situatie doet zich voor bij nieuwe levensreddende behandelingen, zoals bepaalde orgaantransplantaties of de behandeling van kinderen met tot voor kort dodelijke aangeboren afwijkingen. De technische mogelijkheden het leven met dergelijke ingrepen te verlengen roepen onmiddelijk de vraag naar de kwaliteit van dat leven op.

Dit proefschrift gaat over het meten van de kwaliteit van leven als uitkomstmaat van ziekte en zorg. Het begrip 'kwaliteit van leven' in die context is afgegrensd tot 'gezondheid-gerelateerde kwaliteit van leven', samengevat als 'gezondheidstoestand'. Hiermee wordt aangegeven dat factoren die invloed kunnen hebben op de kwaliteit van het leven maar die niet direct verband houden met de gezondheid of met medische behandeling, buiten beschouwing worden gelaten. Gezondheidstoestand wordt in dit proefschrift geoperationaliseerd als het functioneren van de patiënt op fysiek, psychisch en sociaal gebied.
Hoofdstuk 2 geeft een overzicht van het wetenschappelijke veld van meting van de gezondheidstoestand. De plaats van gezondheidstoestand ten opzichte van conventionele klinische variabelen (zoals bloeddruk, bloedonderzoek, ECG, röntgenfoto's en dergelijke) kan worden omschreven als elkaar aanvullend, met elk een ander gebruiksdoel. Klinische parameters worden bij voorbeeld gebruikt om een diagnose te stellen

en, vanwege prognostische waarde, om behandelbeslissingen te onderbouwen. Voor de bepaling van de uiteindelijke effecten van ziekte en interventies is het belangrijk te kijken naar het dagelijks functioneren van patiënten (en natuurlijk ook naar overleving). Dit geldt zowel voor individuen als voor patiënten als groep.

Gezondheidstoestandmeting kan plaats vinden vanuit verschillende perspectieven. Onderscheiden worden het perspectief van een individuele patiënt, bij wie het gaat om de keuze tussen behandelingsalternatieven; het perspectief van groepen patiënten met vergelijkbare ziektekenmerken, waarbij onderzoek van de gezondheidstoestand een indruk geeft van de effecten van een interventie bij dergelijke patiënten in het algemeen; en ten slotte een maatschappelijk perspectief, waarbij gezondheidstoestandinformatie wordt gebruikt ter ondersteuning van beslissingen over de verdeling van (schaarse) middelen, met name binnen de gezondheidszorg. Medische Technology Assessment (MTA) gaat uit van een maatschappelijk perspectief, een klassieke klinische trial van het perspectief van de patiëntengroep. Opgemerkt wordt dat onderzoek naar de effectiviteit van interventies uitsluitend vanuit het perspectief van een patiëntengroep ('no matter what it costs') naar de achtergrond lijkt te verdwijnen.

Het perspectief (anders gezegd: de onderzoeksvraag) bepaalt onder meer de keuze van meetinstrumenten voor de gezondheidstoestand. Meetinstrumenten voor gezondheidstoestand kunnen worden onderverdeeld in 3 hoofdgroepen, te weten generieke, ziektespecifieke en domeinspecifieke instrumenten. Generieke instrumenten zijn, doordat ze niet ziekte- of ziektestadium specifiek zijn, bij uitstek geschikt voor vergelijkingen over de grenzen van diagnoses en interventies heen. Generiek meten van de gezondheidstoestand is noodzakelijk in onderzoek dat plaatsvindt vanuit een maatschappelijk perspectief.

Tenslotte wordt in hoofdstuk 2 het belang van standaardisatie van methoden voor meting van de gezondheidstoestand onderstreept. Zonder standaardisatie zijn onderzoeksresultaten onvergelijkbaar. Het gevolg van deze onvergelijkbaarheid kan zijn dat resultaten van studies van verschillende ziekten niet kunnen worden gebruikt om deze ziekten te ordenen naar de ziektelast die zij veroorzaken, of dat met behulp van MTA's van verschillende interventieprogramma's geen uitspraak kan worden gedaan over de relatieve effectiviteit van deze interventies. Onvergelijkbaarheid heeft ook een suboptimale bijdrage van individuele studies aan de wetenschappelijke 'body of knowledge' tot gevolg. Standaardisatie is essentieel om gebruik te kunnen maken van gezondheidstoestandinformatie bij beleidsvoorbereiding en beleidsbeslissingen. Een landelijk onderzoekersnetwerk, de Werkgroep Onderzoek Gezondheidstoestandmeting, beoogt een bijdrage te leveren aan het proces van standaardisatie.

De hoofdstukken 3 tot en met 7 gaan over beschrijvende gezondheidstoestandmeting. *Hoofdstuk 3* vergelijkt 6 in het Nederlands beschikbare generieke meetinstrumenten, namelijk de SIP, de NHP, de MOS-20, de SF-36, de COOP/WONCA kaarten en het EuroQol instrument. Het concept gezondheidstoestand blijkt gedeeltelijk verschillend geoperationaliseerd in deze instrumenten. Een vergelijking van testeigenschappen (betrouwbaarheid, validiteit) bleek op grond van literatuurgegevens nog niet goed mogelijk, omdat testeigenschappen populatie-afhankelijk zijn en omdat de uitvoering en rapportage van onderzoeken naar testeigenschappen vaak onvergelijkbaar bleken. Geen van de beschikbare instrumenten is echter in alle opzichten superieur aan alle andere. Er is behoefte aan empirisch vergelijkend onderzoek van meetinstrumenten voor gezondheidstoestand.

Hoofdstuk 4 doet verslag van een empirische vergelijking van de NHP en de SIP bij cross-sectionele beschrijving van de gezondheidstoestand van een groep nierdialysepatiënten. De NHP bleek eenvoudiger om in te vullen. Ook waren de NHP-schalen intern consistenter. Met behulp van factoranalyse bleken NHP en SIP efficiënt samen te vatten in twee factoren, namelijk een fysieke en een psychosociale. In de SIP wordt fysiek functioneren benadrukt, terwijl de NHP meer ingaat op psychisch functioneren.

Hoofdstuk 5 bevat een soortgelijke studie. Hier worden de NHP, de SF-36, de COOP/WONCA kaarten en het EuroQol instrument empirisch vergeleken ten aanzien van toepasbaarheid ('feasibility'), betrouwbaarheid, construct validiteit en onderscheidend vermogen tussen klinisch verschillende groepen, bij toepassing in een groep lijders aan migraine en een controle groep. Elk van de 4 instrumenten bleek goed te presteren, met dien verstande dat de multi-item instrumenten (NHP en SF-36) voor gezondheidstoestandmeting in beschrijvende zin meer geschikt bleken dan de classificatie-instrumenten (COOP/WONCA en EuroQol). Sensitiviteit voor veranderingen in gezondheidstoestand over de tijd was geen onderwerp van vergelijking. Ook in dit onderzoek bleken de instrumenten efficiënt samen te vatten in een fysieke en een psychosociale factor.

Hoofdstuk 6 is een voorbeeld van een toepassing van beschrijvende gezondheidstoestandmeting. Het betreft een gecontroleerd dwarsdoorsnede onderzoek naar de invloed van migraine op de gezondheidstoestand, gemeten met de NHP, SF-36, de COOP/WONCA kaarten en het EuroQol instrument. De migrainegroep toonde consistent een iets slechtere gezondheidstoestand dan de controlegroep. Dit verschil kon slechts gedeeltelijk worden toegeschreven aan het meer voorkomen van comorbiditeit (m.n. depressiviteit) in de migrainegroep. Migraine heeft een niet zo sterk, maar onafhankelijk negatief effect op het dagelijks functioneren buiten de aanvallen zelf.

In *hoofdstuk 7* wordt verslag gedaan van een longitudinaal opgezet onderzoek naar de effecten van levertransplantatie op de gezondheidstoestand. Voor hen die de interventie zelf overleven heeft levertransplantatie een zeer gunstig effect op de gezondheidstoestand. Als gezorgd wordt voor adequate informatie van patiënten en behandelaars, en voor uiterste gebruiksvriendelijkheid, blijkt empirische gezondheidstoestandmeting ook bij deze soms zeer zieke patiënten mogelijk.

De meeste bekende instrumenten voor meting van de gezondheidstoestand, zoals de NHP en de SF-36, zijn beschrijvend van aard. Dit betekent dat een score op zo'n instrument de vorm heeft van een profiel: een score is samengesteld uit een score voor bijvoorbeeld fysiek functioneren, een score voor psychisch functioneren en een score voor sociaal functioneren. Voor een aantal toepassingen van gezondheidstoestandmeting, zoals in kosten-utiliteitsanalyse en in volksgezondheidsmodellen, is het nodig profielscores samen te vatten in één getal, teneinde effecten op de duur en de kwaliteit van de overleving te kunnen combineren. Een waarderingsprocedure voor gezondheidstoestanden is een methode om profielscores voor gezondheidstoestand samen te vatten in één getal.

Hoofdstuk 8 beoogt een overzicht te geven van de stand van de wetenschap op het gebied van het waarderen van gezondheidstoestanden. De op dit moment gangbare 3-staps procedure wordt toegelicht. Beschrijvingen van de gezondheidstoestand van patiënten (verkregen in stap 1) worden in de tweede stap gewaardeerd, waarna in de derde stap de waarderingen worden gecombineerd met overlevingsduur. De uitkomsten worden uitgedrukt in bijvoorbeeld Quality Adjusted Life-years (QALY's). Belangrijke keuzen in de procedure worden besproken. Zo wordt betoogd dat de keuze van het groep die de waarderingen geeft bepaald wordt door het perspectief van het onderzoek. In het geval van een maatschappelijk perspectief dienen de waarderingen het maatschappelijk gezichtspunt te representeren; dit kan worden geoperationaliseerd door de waarderingen te laten uitspreken door een representatieve steekproef uit de algemene populatie (waarin dus ook patiënten vertegenwoordigd zijn). Een andere belangrijke keuze betreft de waarderingsmethode. Een visueel analoge schaal biedt praktische voordelen. Time trade-off is mogelijk te prefereren uit een oogpunt van validiteit. Directe transformatie van resultaten van de ene methode naar de andere is in principe mogelijk als de rangorde van de gezondheidstoe-standen dezelfde is.

In de operationalisatie van het QALY-concept in de beschreven 3-staps procedure worden de uitkomsten van een ziekte of een interventie in feite opgeslitst. De uitkomstruimte kan worden voorgesteld als een boom met verschillende takken, die elk worden gekenmerkt door een duur, een sequentie van gezondheidstoestanden en een frequentie van voorkomen. De huidige QALY-benadering gaat uit van onder andere de volgende 2 aannames, die vermoedelijk zelden geheel juist zijn. Ten eerste wordt verondersteld dat de waardering voor een gezondheidstoestand niet afhangt van wat eraan voorafgaat (historie) en wat volgt (prognose). Theoretisch zou het waarderen van gezondheidstoestandsequenties ('belopen') realistischer zijn. Ten tweede wordt verondersteld dat de waardering voor een gezondheidstoestand onafhankelijk is van de duur ervan.

Indien onderzoekers en gebruikers (beleidsmakers) zich bewust zijn van de beperkingen én de kracht van het QALY-concept is het echter gerechtvaardigd door te gaan met het ontwikkelen en verfijnen van de operationalisatie ervan. Er is bovendien nog geen alternatief dat aantoonbaar beter is.

De hoofdstukken 9 en 10 betreffen empirische waarderingsstudies in de algemene populatie. Beiden waren deel van het onderzoeksprogramma van de EuroQol groep, een Europees netwerk van onderzoekers dat tot doel heeft om door middel van een gestandaardiseerde empirische aanpak (met behulp van het EuroQol instrument) een methodologische bijdrage te leveren op het gebied van waarderingen van gezondheidstoestanden. *Hoofdstuk 9* beschrijft een onderzoek naar de haalbaarheid van het verzamelen van waarderingen in de algemene populatie met een postenquete. De resultaten waren veelbelovend. Er waren echter een aanzienlijke non-respons en nietgeslaagde respons (van respondenten voor wie de vragenlijst blijkbaar te moeilijk was; met name ouderen en lager opgeleiden). Dit was aanleiding een non-respons onderzoek te doen, dat wordt beschreven in *hoofdstuk 10.* Hoewel de effecten van selectieve nonrespons en niet-geslaagde respons lastig te evalueren bleken, lijkt de relevantie van dergelijke effecten in verband met het gebruik van resultaten bij beleidsbeslissingen gering. Relevante effecten van achtergrondvariabelen (leeftijd, opleiding) op de waarderingen werden niet aangetoond.

De belangrijkste conclusies van dit proefschrift komen aan de orde in *Hoofdstuk 11.* Gezondheidstoestand is, naast overlevingsduur en klinische parameters, de derde pijler van medisch evaluatieonderzoek. Voor beschrijvende gezondheidstoestandmeting zijn de methoden zodanig gerijpt dat dit een standaardonderdeel van elk evaluatie onderzoek moet uitmaken: niet het opnemen van gezondheidstoestandmeting in de onderzoeksopzet, maar juist het eruit laten ervan zou beargumenteerd moeten worden. Standaardisatie van methoden is noodzakelijk ten behoeve van de vergelijkbaarheid en daarmee de bruikbaarheid van onderzoeksresultaten ten behoeve van beleid. In het bereiken van het gewenste niveau van standaardisatie is een belangrijke rol weggelegd voor de Werkgroep Onderzoek Gezondheidstoestandsmeting én voor de gebruikers van gezondheidstoestandinformatie, bij voorbeeld de overheid.

Voor het waarderen van gezondheidstoestand is de situatie anders. Op dit gebied wachten nog belangrijke vragen, bijvoorbeeld naar de rol van 'tijd' en 'beloop', en het bestaan van subgroepen in de populatie met afwijkende waarderingsprofielen, op een empirisch antwoord. Routinematig gebruik van waarderingsuitkomsten vereist daarom op dit moment nog enig voorbehoud. Systematisch methodologisch onderzoek op terrein van de waardering van uitkomsten verdient krachtige steun.

List of publications

Agt HME van, Essink-Bot ML, Krabbe PFM, Bonsel GJ. *Test-retest reliability of EuroQol valuations on health states.* Soc Sci Med 1994;11:1537-1544.

Bonsel GJ, Habbema JDF, Bot ML, Veer F van 't, Charro FT de, Maas PJ van der. *Een 'technology assessment' van levertransplantatie; een onderzoek naar het Groningse levertransplantatieprogramma van 1977 - 1987.* Ned Tijdschr Geneeskd 1989;133: 1406-1414.

Bonsel GJ, Essink-Bot ML, Charro FT de, Maas PJ van der, Habbema JDF. *Orthotopic liver transplantation in the Netherlands. The results and impact of a medical technology assessment.* Health Policy 1990;16;147-161.

Bonsel GJ, Klompmaker IJ, Essink-Bot ML, Habbema JDF, Slooff MJH. *Cost-effectiveness analysis of the Dutch liver transplantation programme.* Transplantation Proceedings 1990;22:1481-1484.

Bonsel GJ, Essink-Bot ML, Klompmaker IJ, Slooff MJH. *Assessment of quality of life before and following liver transplantation: first results.* Transplantation 1992;53:796-800.

Essink-Bot ML, Bonsel GJ, Maas PJ van der. *Valuation of health states by the general public: feasibility of a measurement procedure.* Social Science & Medicine 1990;31: 1201-1206.

Essink-Bot ML, Agt HME van, Bonsel GJ. *NHP of SIP - een vergelijkend onderzoek onder chronisch zieken.* Tijdschrift Sociale Gezondheidszorg 1992;3:152-159.

Essink-Bot ML, Stouthard MEA, Bonsel GJ. *Generalizability of valuations on health states collected with the EuroQol-questionnaire.* Health Economics 1993;2:237-246.

Essink-Bot ML, Bonsel GJ. *Letter to the Editor (regarding papers on the Quality of Life in Depression Scale by McKenna and Hunt).* Health Policy 1993;23:265-266.

Essink-Bot ML. *De Werkgroep Onderzoek Gezondheidstoestandmeting: een bijdrage aan standaardisatie van onderzoek naar met gezondheid gerelateerde kwaliteit van leven (Brief aan de Redactie).* Ned Tijdsch Geneeskd 1994;138:1484-1486.

Essink-Bot ML, Bonsel GJ. *Naar standaardisatie van het instrumentarium voor het meten van de gezondheidstoestand.* Huisarts & Wetenschap 1995;38(3):117-122.

Essink-Bot ML, Royen L van, Krabbe PFM, Bonsel GJ, Rutten FFH. *The impact of migraine on health status.* Headache 1995;35:200-206.

Essink-Bot ML, Krabbe PFM, Agt HME van, Bonsel GJ. *NHP or SIP: a comparative study in renal insufficiency associated anemia.* In press in Quality of Life Research.

Mannes GPM, Essink-Bot ML, Koëter GH, Bosscher D, Boer WJ de, Bij W van der, Vergert EM Ten. *Lung transplantation and quality of life.* In press in European Respiratory Journal.

Roijen L van, Essink-Bot ML, Koopmanschap MA, Michel BC, Rutten FFH. *A society's perspective on the burden of migraine in the Netherlands.* Pharmaco-Economics 1995;7(2):170-179.

Roijen L van, Essink-Bot ML, Koopmanschap MA, Bonsel GJ, Rutten FFH. *Labour and health status in economic evaluation of health care; the Health & Labour Questionnaire.* In press in International Journal for Technology Assessment in Health Care.

Dankwoord

Het schrijven van een welgemeend dankwoord zonder al te veel cliché's is misschien wel lastiger dan de produktie van een wetenschappelijke tekst. Natuurlijk is dit proefschrift met hulp van velen tot stand gekomen. Ik wil de gelegenheid gebruiken om enkelen van hen hier speciaal te noemen.

Dat is in de eerste plaats mijn promotor, Prof.Dr. P.J. van der Maas. Beste Paul, ik wil je danken voor het feit dat je me in de gelegenheid hebt gesteld mijn proefschrift te schrijven. Dit 'in de gelegenheid stellen' slaat natuurlijk op de inhoudelijke begeleiding en de organisatorische voorwaarden. Het slaat ook op je flexibiliteit en begrip als baas, die naar mijn mening veel hebben bijgedragen aan het feit dat ik een wetenschappelijke carrière tot nu toe heb kunnen combineren met een thuissituatie die veel energie vraagt.

Met mijn co-promotor Dr. G.J. Bonsel heb ik vele jaren samengewerkt in de onderzoeksprojecten waarop dit proefschrift is gebaseerd. Beste Gouke, laat ik volstaan met te zeggen dat ik zeer veel van je heb geleerd. Woorden schieten tekort - alleen, 'grijs' was het nooit.

Een belangrijk deel van mijn wetenschappelijke activiteiten heeft zich in iMTA-verband afgespeeld. Prof.Dr. F.F.H. Rutten, beste Frans, en natuurlijk alle iMTA collega's met wie ik heb samengewerkt, ik denk met veel plezier aan het werk voor iMTA terug. Dat geldt in het bijzonder voor drs. L. van Roijen. Beste Leona, het migraineproject, de ontwikkeling van de 'Ziekte & Werk' vragenlijst - het waren succesvolle en leuke ondernemingen. Ik hoop dat er ondanks de fysieke scheiding tussen iMTA en iMGZ mogelijkheden zullen blijven bestaan om in deze lijn verder samen te werken.

De leden van de Werkgroep Onderzoek Gezondheidstoestandmeting dank ik voor hun medewerking en inzet, zonder welke een netwerk van onderzoekers niet kan functioneren. Prof.Dr. J. Passchier, beste Jan, de Werkgroep heeft in jou een goede voorzitter. Als de harmonie waarin het 'besturen' van de Werkgroep tot nu toe is verlopen voorspellende waarde heeft voor haar toekomst, heb ik goede hoop. Daarnaast dienen zich ook gezamenlijke projecten voor iMGZ en het Instituut voor Medische Psychologie en Psychotherapie aan buiten Werkgroepverband. Prof.Dr. J.C.J.M. de Haes, beste Hanneke, jou wil ik in het bijzonder bedanken voor je bijdrage aan hoofdstuk 2.

Mijn naaste collega's in de loop der jaren waren of zijn drs. H.M.E. van Agt, drs. P.F.M. Krabbe, drs. E.N.T.M. van Lin, Dr. M.E.A. Stouthard, en de student-assistenten M. Palmen en A. Bandel. Beste Heleen, Paul, Emile, Marlies, Maurice, Arjan: bedankt voor al het werk, de goede ideeën, de kritische reflectie en, wederom, de prettige samenwerking.

Dr. E.M. ten Vergert, beste Els, drs. K. Stronks, beste Karien, ik wil jullie bedanken voor het kritisch commentaar op delen van het proefschrift die niet als wetenschappelijk artikel gepubliceerd zijn.
Inhoudelijk is dit proefschrift mijn verantwoordelijkheid; Mw. I. Philips heeft de rest gedaan. Beste Ilse, als de inhoud even mooi is als het uiterlijk hebben we samen goed werk geleverd. Daarnaast dank ik Paul Krabbe voor de lay-out van de figuren.
I am grateful to Ms. R. Rabin, dear Rosalind, whose knowledge of the topic was most helpful in facilitating the accurate correction of the English.

Ten slotte de mensen zonder wie er echt niets van terecht zou zijn gekomen. Dat zijn de directie, leidsters en andere medewerkers van kinderdagverblijf 't Kinderparadijs te Rotterdam. Anita, Astrid, Inge, Jacqueline, Jeannette, Jolanda, Judith, Karin, Margriet, Maria, Nuria, Sarinke, Simone, Urmi, en alle anderen - de liefde en het verantwoordelijkheidsgevoel waarmee jullie je werk doen zijn echt bijzonder.

Lieve Jan, Bas en Erik. Waarom maakt mama eigenlijk boeken en kranten in de witte flat? Jullie relativeren ongewild een hoop 'moeten' en daar ben ik jùllie dankbaar voor. De gezondheidstoestand en de kwaliteit van leven van Erik zijn een bijzondere motivatie voor het werk juist op het terrein van gezondheidstoestandmeting.
Lieve Rob: ik hoop dat we ons 'teamwork' in de toekomst op een iets minder hectische manier kunnen voortzetten.

Curriculum Vitae

Marie-Louise Bot werd geboren op 10 september 1960 te Utrecht. In 1978 haalde zij het einddiploma gymnasium-ß (Corderius College, Amersfoort). In 1985 sloot zij haar studie geneeskunde aan de Rijksuniversiteit Utrecht af. Gedurende 1985 - 1987 was zij werkzaam als arts-assistent Interne Geneeskunde en Cardiologie in resp. het St.Jozef Ziekenhuis te Kerkrade en het St.Clara Ziekenhuis te Rotterdam. Sinds 1987 is zij werkzaam als onderzoeker bij het Instituut Maatschappelijke Gezondheidszorg (hoofd: Prof.Dr. P.J. van der Maas), Erasmus Universiteit Rotterdam. In 1993 werd zij geregistreerd als epidemioloog. De opleiding tot sociaal-geneeskundige (zonder takaanduiding) werd afgesloten in 1994.

Belangrijke onderzoeksprojecten waaraan Marie-Louise Essink-Bot heeft meegewerkt zijn o.a. de technology assessments van hart-, lever- en longtransplantatie (gefinancierd door de Ziekenfondsraad); het programma 'Standaardisering van Medische Technology Assessment' (gefinancierd door het Stimuleringsprogramma Gezondheidsonderzoek), diverse ontwikkelingsgeneeskunde projecten en een onderzoek naar migraine (gefinancierd door Glaxo BV). Vaak werd nauw samengewerkt met het Instituut voor Medische Technology Assessment (hoofd: Prof.Dr. F.F.H. Rutten) van de Erasmus Universiteit. Lopende onderzoeken betreffen o.a. de kwaliteit van leven effecten van screening op prostaatkanker, de kwaliteit van leven bij leverziekten, en vergelijkend onderzoek van meetinstrumenten voor kwaliteit van leven / gezondheidstoestand. Marie-Louise Essink-Bot is wetenschappelijk secretaris van de landelijke Werkgroep Onderzoek Gezondheidstoestandmeting en actief lid van de internationale EuroQol Group.

Marie-Louise Essink-Bot is getrouwd met Rob Essink. Zij hebben drie kinderen, Jan (1991) en de tweeling Bas en Erik (1993).