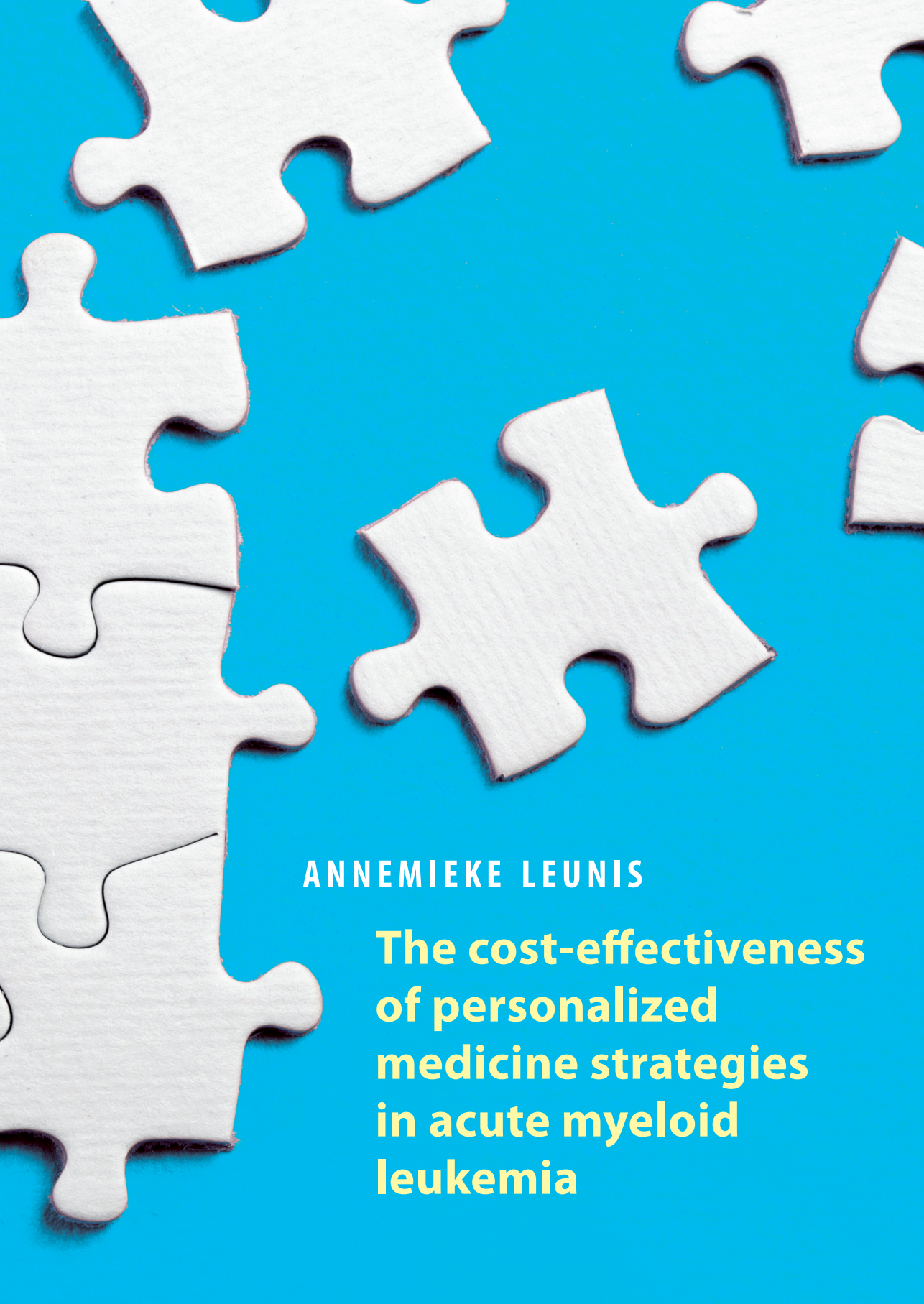


ANNEMIEKE LEUNIS

The cost-effectiveness of personalized medicine strategies in acute myeloid leukemia

The cost-effectiveness of personalized medicine strategies in acute myeloid leukemia

Annemieke Leunis

Layout and printing: Optima Grafische Communicatie, Rotterdam, The Netherlands

ISBN: 978-94-6169-679-3

Funding
The studies in this thesis were supported by the Center for Translational Molecular Medicine (CTMM), project BioCHIP (grant 030-102), the Netherlands Organization for Health Research and Development (ZonMw) and the Dutch Society of Hematology.

The cost-effectiveness of personalized medicine strategies in acute myeloid leukemia

De kosteneffectiviteit van geïndividualiseerde behandelstrategieën in acute myeloïde leukemie

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

donderdag 18 juni 2015 om 13.30 uur

door
Annemieke Leunis
geboren te Strijen

PROMOTIECOMMISSIE:

Promotoren:
Prof.dr. C.A. Uyl-de Groot
Prof.dr. B. Löwenberg

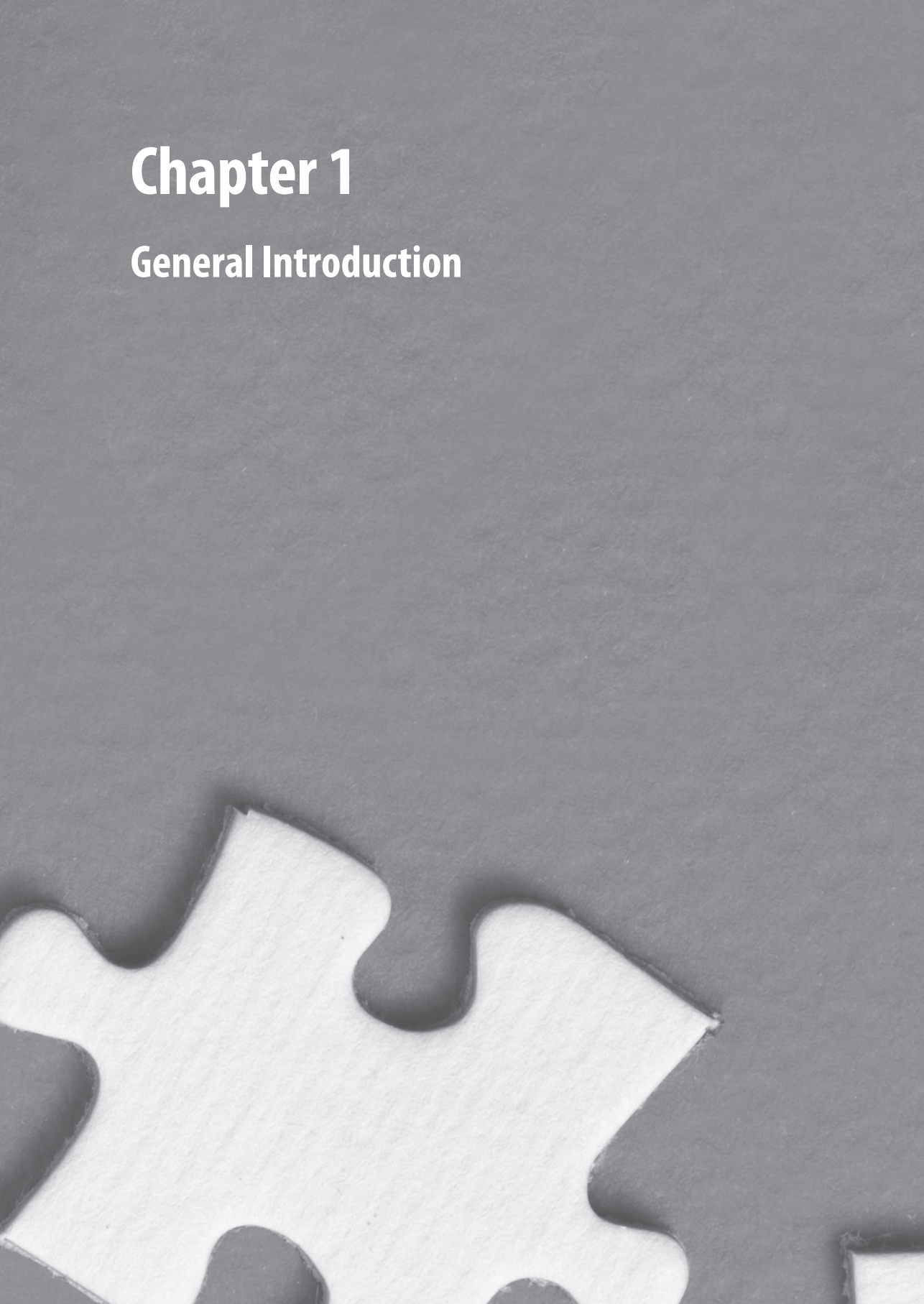
Overige leden:
Prof.dr. W.B.F. Brouwer
Prof.dr. J.J. van Busschbach
Prof.dr. N.M.A. Blijlevens

Copromotor:
Dr. W.K. Redekop

CONTENTS

Chapter 1

General Introduction



BACKGROUND

Cancer is one of the most important health problems in the developing world with an estimated incidence of 3.5 million per year and 1.75 million deaths in Europe in 2012 (1). Thereby cancer attributes to about 28% of all deaths and is the second leading cause of death in Europe (2). Although cancer is one of the most important causes of death nowadays, survival after cancer has been improved over time (3-5). The improvements in survival are caused by earlier detection of cancer and new treatment developments. The earlier detection enables a more adequate treatment with higher chances of cure. However, the earlier detection can also artificially improve survival as the time to death will increase if the diagnosis has been set earlier (4). The new treatment options include new technologies and pharmaceuticals.

Due to the large incidence of cancer and the high costs of the new technologies and pharmaceuticals to treat cancer, the economic burden of cancer is substantial (6). The high mortality and morbidity of cancer also poses a large burden on society in terms of productivity loss and informal care costs. It is expected that the costs of cancer will increase in the future (7). This increase will be partly caused by an increase in cancer incidence due to the aging of the population as cancer occurs more frequently in older patients. Furthermore, new developments in diagnosis and treating cancer are often associated with high costs. A balance needs to be found between the reduction in mortality and morbidity and the rising health care costs due to the new technologies.

One of the new technology developments in the field of cancer is personalized medicine. Personalized medicine is a term that is used for medicine which is targeted to a specific patient group. The main reasons for the development of personalized medicine is that a one-size-fits-all treatment approach does not always lead to the most desirable outcomes as not all patients respond to the treatment. Response rates of cancer drugs approved 1995 and 2005 ranged between 10 and 80% (8,9). Currently, new techniques enable a better identification of responders and non-responders before the start of the treatment (10). If treatment will be restricted to the responders only, the effectiveness of the treatment will increase. Furthermore, the costs might decrease as fewer patients receive the expensive treatment and fewer side-effects are experienced because non-responders are no longer exposed to an ineffective and possible toxic treatment.

PERSONALIZED MEDICINE

The term personalized medicine is often used to characterize treatment tailored according to genetic information (11). However, genetic information is not the only information source to identify patients who should receive a specific treatment or not. Other relevant characteristics may include age, comorbidities, performance status and regular blood tests results (12). Intensive cancer treatment is often restricted to younger patients with a good performance status and few comorbidities (13-16). It has also been shown that patients with comorbidities have a poorer outcome, irrespective of age and stage of the disease (17). As other information than genetic test results can also be used to tailor treatment, the following definition of personalized medicine is used in this thesis: 'the use of combined knowledge (genetics, or otherwise) about a person to predict disease susceptibility, disease prognosis or treatment response and thereby improve that person's health' (12).

The above mentioned definition also distinguishes different applications of personalized medicine. Table 1.1 shows three different applications of personalized medicine as well as examples of genetic and non-genetic information used to personalize treatment.

The first application is the use of individual information to predict disease susceptibility. This information is especially useful if preventive measures are available to avoid the development of the disease, like mastectomy for people with a high risk of breast cancer (24). The second application is the use of individual characteristics to identify the prognosis of the patient. This information might have important consequences for the treatment choice if patients with a good prognosis are treated differently than patients with a poor prognosis. Finally, individual information can also be used to predict treatment response, development of side effects and adequate dosing of drugs to guide treatment decisions.

Although personalized medicine is not only related to the use of genetic information, the knowledge about the human genome has increased the possibilities to individualize treatment. Many studies have found genetic heterogeneity in the known cancer types and associations between genetic markers and prognosis (25-27). It is therefore expected that many new personalized medicine strategies will be developed in the future.

The effectiveness and the cost-effectiveness of new strategies needs to be evaluated before these strategies can be implemented in daily practice. The rationale for cost-effectiveness analyses is that resources are scarce and motivated choices are requirered for investing these scarce resource in new health care technologies. As it is expected that the health care expenditures will rise in the future (28), it becomes even more important to make these motivated choices.

Table 1.1 Different applications and examples of personalized medicine

Application of personalized medicine	Example non-genetic information	Example genetic information
Predict disease susceptibility	Healthy lifestyle factors (healthy weight, high physical activity, non-smoking, limited alcohol consumption and a healthy diet) are associated with a lower lower incidence of colorectal cancer in Europe (18).	BRAC mutations in breast cancer patients (19).
Identify prognosis of the patient	Prognostic indices for brain metastases. These indices include the following parameters: age, performance status, presence of extracranial or brain metastases (20).	Risk-stratified treatment for acute myeloid leukemia based upon cytogenetic and molecular abnormalities (21).
Predict treatment response or side effects	Geriatric factors (cognitive functioning, dependence, depression) are predictive of severe toxicity or unexpected hospitalization in patients with metastatic colorectal cancer (22).	Erlotinib and gefitinib for non-small lung cancer patients with mutations in the gene encoding the epidermal growth factor receptor (EGFR) (23).

COST-EFFECTIVENESS ANALYSIS

In cost-effectiveness analyses, both the costs and effects of a new product are compared with the costs and effects of current products (29). The cost-effectiveness of a product is reported as the incremental cost-effectiveness ratio (ICER), which can be calculated with the following formula:

$$ICER = \frac{Costs\ New - Costs\ Old}{Effect\ New - Effects\ Old}$$

The effects in cost-effectiveness analyses are often measured in quality-adjusted life years (QALYs). A QALY is measure which combines mortality and health-related quality of life in one measure. Within the QALY concept, health-related quality of life is expressed as a utility of being in a specific health state. QALYs are calculated by multiplying the life years in a specific health state with the quality of life utility associated with that health state. Perfect health is represented by a utility value of 1 and death with a utility value of 0.

Many different questionnaires are currently available to measure health-related quality of life. Guidelines for cost-effectiveness analyses recommend the use of a generic quality-of-life questionnaire, the EQ-5D, for use in cost-effectiveness analyses unless there is evidence that the EQ-5D has a poor performance in the selected patient population (30). The main reason for this recommendation is to guarantee the comparability of cost-effectiveness analyses between diseases as it has been shown that utility values differ between questionnaires (31).

If it has been shown that the EQ-5D has a poor performance in the selected population, the utility values needs to be derived from a more valid questionnaire. This questionnaire can either be another generic questionnaire or a disease-specific questionnaire. However, are often not developed to calculate quality of life utilities. Preference-based utility values need to be estimated for these questionnaires. In general, two different methods exit to estimate utility values for disease-specific questionnaires. The first method is the prediction of generic quality of life utilities based upon the answers given to the disease-specific questionnaire (32). This method is called 'mapping'. The second method is the direct valuation of health states from the disease-specific questionnaire (33). The direct valuation method is recommended if it has been shown that the EQ-5D has a poor performance in the selected population. The mapping method is especially useful if quality of life has only been measured with a disease-specific questionnaire although evidence is available for the validity of the EQ-5D in the patient population under study (34).

Cost-effectiveness analyses are often performed from a life-time perspective including all costs and effects in the remaining period of life (29). With respect to personalized medicine strategies, it means that not only the costs of testing for specific patient characteristics should be measured, but also the costs of the subsequent treatment choices. Furthermore, it is recommend that cost-effectiveness analyses are performed from a societal perspective and thereby include all costs and effects for the society as a whole (29). This means that not only medical, but also non-medical costs should be included. Medical costs include the costs of medication, hospital visits, laboratory tests and radiology. Non-medical costs include productivity costs, which are costs of absenteeism at work or reduced productivity at work due to the disease, traveling costs and informal care costs (35,36).

Cost-effectiveness analyses can be performed alongside a clinical trial (piggy-back economic evaluation). Information about costs, quality of life and survival will be collected during the clinical trial. However, a clinical trial has often a limited time period (about 3-5 years), while a lifetime perspective if preferred for cost-effectiveness analyses. Modeling methods are often used to extend the results of the trial to a longer time period (29).

The extrapolation of the trial period is not the only role for modeling in cost-effectiveness analyses. Modeling is extremely useful if data has been derived from different sources, like genomic, clinical and epidemiological studies (37,38). The information from all these sources can be synthesized and combined in the model to assess cost-effectiveness of new strategies (29). Furthermore, modeling can also be used for the planning of future studies, including the type of data collection and the focus in terms of target population. Once modeling studies show a substantial impact of changes in certain input parameters on the ICER (39), it is important that more information about these parameters will

be collected in future studies. As these studies will come with additional costs, modeling can also be used to identify the value of additional research (40). A modeling study might also identify areas with the largest medical needs. Preferably, new studies should focus on those areas. Furthermore, modeling can also be used early in the development process to evaluate whether it is worthwhile to continue the development of a product (41). This early evaluation includes an assessment of the required effectiveness of a new product given a certain price and cost-effectiveness threshold. Furthermore, the maximum possible price of future products can be estimated. Continuation of product development is only useful if that price will cover all development and production costs.

It is expected that systematic use of cost-effectiveness analyses in the development and assessment of personalized medicine strategies in cancer will improve the implementation of these new strategies in daily practice. However, it needs to be evaluated how cost-effectiveness analyses can be systematically applied during the development process of new technologies. This thesis aims to answer this question by a thorough evaluation in a specific type of cancer with a high potential for personalized medicine, namely acute myeloid leukemia (AML). Besides the high potential of personalized medicine in the field of AML, the current prognosis of AML is still very poor. Furthermore, current treatments for AML are expensive and have a large impact on the health-related quality of life. Therefore, new treatment options are needed to improve survival and health-related quality life and reduce costs in patients with AML.

ACUTE MYELOID LEUKEMIA

Acute myeloid leukemia (AML) is a specific type of leukemia which is characterized by a proliferation of immature myeloid cells (blasts) in the bone marrow. The proliferation of blasts reduces the development of normal blood cells which lead to increased risk of bleedings and infections and make patients feel tired and weak (42). The rapid increase of blasts in patients with AML requires immediate treatment to reduce the number of blasts and restore the normal blood function (26).

The large potential of personalized medicine in the field of AML is caused by the heterogeneous nature of the disease (26). Many different subtypes of AML are identified based upon cytogenetic and molecular abnormalities. Some of these subtypes are already defined as a distinct entity in the most recent classification of the World Health Organization (WHO) (43). The prognosis differs between patients with different cytogenetic and molecular abnormalities (44-47). Due to the differences in pathogenesis and prognosis of the distinct subtypes, treatment already differs between subgroups. One specific type

of AML, acute promyelocytic leukemia (APL) is already treated with a targeted treatment regimen (48). This targeted treatment regimen has dramatically improved the prognosis of APL (49). For all other patients, a risk-stratified treatment is applied with less intensive treatments for patients with a low risk of relapse (26,50). However, research is ongoing to find targeted treatments for other subgroups of AML (51,52).

In general, treatment for AML consists of chemotherapy and/or hematopoietic stem cell transplantation (26). All treatments are very intensive and require hospitalization to prevent the development of infections. The long hospitalization is associated with high treatment costs. Another negative consequence of the intensive treatments is the impact of the treatments on health-related quality of life. AML patients report frequently problems with physical, psychological and emotional and sexual functioning (53). Furthermore, it has been shown that the health-related quality of life is more reduced in patients receiving more intensive treatments like hematopoietic stem cell transplantations (54-56). However, not all patients can adequately be treated with current treatment options resulting in a poor overall prognosis. Only 20% of the patients with AML is alive 5 years after the diagnosis (57). The prognosis is better for younger patients; the 5-year overall survival is 55% and 9% in patients aged 18-44 and 65-74 years, respectively. An important factor for the poor prognosis in older patients is the inability of older patients to receive the intensive treatments. New treatments are needed to adequately treat elderly patients, improve current cure rates and health-related quality of life and reduce the treatment costs. As some of the patients with AML can already be adequately treated with current treatment options, it is expected that the new treatments will be focused on specific subgroups.

Personalized medicine will always reduce the size of the population under study. This is more problematic in AML compared to other cancers, due to the low incidence of AML. The incidence of AML is 3-4 per 100,000 compared to a breast cancer incidence of 94 per 100,000 (1,58). Due to the low incidence of AML and its subgroups, international collaborations are needed for sufficiently powered trials to detect a significant difference in survival or response (52). It is important that future studies are set up as efficiently as possible to obtain maximum health gain with a minimum amount of resources.

RESEARCH QUESTIONS

In order to assess the cost-effectiveness of personalized medicine strategies in AML and provide methodological guidance for evaluations in other disease areas, the following research questions have been defined:

- What are the costs of acute myeloid leukemia treatment in the Netherlands?
- What is the quality of life of acute myeloid leukemia patients in the Netherlands?
- Which HRQOL instrument should be used to measure HRQOL in acute leukemia patients for the purpose of an economic evaluation?
- What is the cost-effectiveness of personalized medicine strategies for AML patients?
- Which recommendations can be given with respect to future developments in the field of AML?
- Which recommendations can be given with respect to the methodology for cost-effectiveness analyses of personalized medicine strategies in other disease areas?

OUTLINE OF THE THESIS

This thesis consists of two parts. The first part of the thesis (chapter 2-6) focuses on studies performed to obtain inputs in terms of costs and health-related quality of life for use in cost-effectiveness analyses in the field of AML due to limited information in the literature. The second part (chapter 7-8) describes the development of a decision model for AML and the application of the model in cost-effectiveness analyses.

Chapter 2 reports the results of a microcosting study performed in three hospitals in the Netherlands. Resource use was collected for all patients diagnosed with de novo or secondary AML and started with induction treatment in 2008 or 2009. The total treatment of AML was distinguished in three phases: induction treatment, post-remission treatment and follow-up.

Chapter 3 and 4 describe the results of the two studies performed to estimate utility values for the cancer-specific quality of life questionnaire EORTC QLQ-C30. The mapping study of the QLQ-C30 onto the EQ-5D is described in chapter 3. EQ-5D utilities were predicted according to the answers given to the QLQ-C30. This prediction algorithm was developed in patients with multiple myeloma. Multiple myeloma patients have filled in both the EQ-5D and the QLQ-C30 at different points in time. For the development, the results of all time points were pooled to increase the sample size. The resulted algorithm was validated in patients with non-Hodgkin lymphoma.

The results of the direct valuation study are described in chapter 4. First, the QLQ-C30 has been reduced to an 8-item instrument, the QLQ-PBM, according to expert opinion, psychometric criteria and fit to the Rasch model. A selection of the possible health states were valued by the general public using the time-trade-off method. The utility values of the other health states were estimated with regression methods. Both the mapping

and direct valuation study were not restricted to the QLQ-C30, but utility values were also estimated for two other disease-specific questionnaires, the HAQ for rheumatoid arthritis patients and the MSIS-29 for multiple sclerosis patients.

Chapter 5 shows the results of a cross-sectional quality of life study in AML survivors of one academic hospital in the Netherlands. Both the results of the new version of the EQ-5D, the EQ-5D-5L, and the QLQ-C30 were reported. The study compared the quality of life of AML survivors with that of the general population to assess the impact of AML and its treatments on quality of life. Furthermore, the study provided a first assessment of factors associated with quality of life in AML.

Chapter 6 reports the results of the study which assess the feasibility, validity and reliability of the EQ-5D-5L and the QLQ-PBM in acute leukemia patients. The study was based upon data of the cross-sectional quality of life study described in chapter 5, but extended with data from patients with acute lymphoblastic leukemia. This chapter provided some guidance regarding the quality of life instrument that should be used in cost-effectiveness analyses in the field of AML.

Chapter 7 describes the development and validation of a decision model representing the full disease course of AML. A discrete-event-simulation model was developed to incorporate relevant patient, disease and treatment characteristics. The validation of the model consisted of face validation by clinical expert and an internal and external validation. The internal validation included a comparison of the clinical outcomes in the model with the original data. For the external validation, the clinical outcomes were compared with the published results of clinical trials in other countries.

The cost-effectiveness of a revised risk group classification is described in chapter 8. The validated model is used to perform this cost-effectiveness analysis. The costs and quality of life estimates were derived from the studies performed in the Netherlands. Both univariate and probabilistic sensitivity analyses were performed to assess the impact of uncertainty in the input parameters on the cost-effectiveness of the new risk group stratification.

The main results of this thesis are described in chapter 9. This chapter provides practical recommendation of future research in the field of AML. Furthermore, the application of the methods in this thesis for assessments of personalized medicine strategies in other disease areas will also be addressed in the general discussion.

Chapter 2

The costs of initial treatment for patients with acute myeloid leukemia in the Netherlands

With Hedwig M. Blommestein, Peter C. Huijgens, Nicole M.A. Blijlevens, Mojca Jongen-Lavrencic and Carin A. Uyl-de Groot

Published in Leukemia Research 37(3): 245-50

ABSTRACT

The aim of this study was to calculate the costs of the current initial treatment of acute myeloid leukemia. Resource use was collected for 202 patients who started with intensive chemotherapy in 2008 or 2009. The costs of the first induction course were significantly higher than the costs of the second induction course. Allogeneic transplantation from a matched unrelated donor was significantly more expensive than the other consolidation treatments. In-hospital stay was the major cost driver in the treatment of AML. Research regarding possibilities of achieving the same or better health outcome with lower costs is warranted.

INTRODUCTION

Acute myeloid leukemia (AML) is an aggressive disease which requires intensive treatment. Treatment of AML generally consists of several induction chemotherapy courses to induce complete remission (CR). Induction treatment is followed by consolidation treatment consisting of high-dose chemotherapy or autologous or allogeneic hematopoietic stem cell transplantation (HSCT) for patients younger than 60-65 years of age. Stem cell sources of allogeneic HSCT are threefold: HLA-identical sibling, matched unrelated donor (MUD) or umbilical cord blood (UCB) (26). The choice of consolidation treatment depends on the patient's risk of relapse and treatment-related mortality (59).

Insight into the treatment costs is an essential requirement for adequate reimbursement of treatment. In addition, cost calculations are required as input for cost-effectiveness analyses of (new) treatments. The cost-effectiveness factor will become increasingly important due to rising health care expenditures in Western countries (60). A few studies in the 90s and early 2000s have calculated the total costs of AML treatment (61-63). However, treatment strategies have changed dramatically (64-67), and an update of the treatment costs is therefore essential. The aims of this study were to gain insight into the current treatment costs and the different cost components of the total treatment costs.

MATERIALS AND METHODS

Patients

All adult patients diagnosed with de novo primary or secondary AML who started with induction chemotherapy in 2008 or 2009 in three university hospitals in the Netherlands were included in this study. Patients with acute promyelocytic leukemia (APL) were excluded because the number of patients was small and these patients were treated differently. Data were collected from diagnosis until relapse, death or last day of registration (June 2011).

Treatment

The initial treatment of AML was distinguished in three treatment phases: induction treatment, consolidation treatment and follow-up. Induction treatment started at the day of diagnosis. Two different treatment protocols were used based on the patient's age. Younger patients (less than 65 years) received induction and consolidation treatment. Induction treatment consisted of cytarabine and idarubicin in the first course and cytarabine and amsacrine in the second course. Five different consolidation treatments were administered: high-dose chemotherapy, autologous HSCT, allogeneic HSCT from a

sibling donor, allogeneic HSCT from a MUD and UCB transplantation. The choice of treatment depended on the patient's risk of relapse, performance status and the availability of an HLA-identical donor. Older patients received induction treatment consisting of cytarabine and daunorubicin in the first course and cytarabine in the second course. In both age groups, the second induction course was administered to all patients, irrespective of achievement of a CR after the first course. The second induction course started the day after discharge for the first course. In case of a continuous hospitalization, the second course started on the day that cytarabine was given as part of the second course. Consolidation treatment started on the day after discharge for the second induction course. Follow-up started 42 days after induction treatment in older patients and 42 days after consolidation treatment in younger patients. Follow-up was set at 1 year or ended at the date of relapse or death.

Cost calculation

The microcosting method was used to calculate the direct hospital treatment costs of AML. All medical resource use related to the treatment of AML and its complications was collected and multiplied by the unit cost of each resource use.

Resource use was derived from electronic patient charts and hospital information systems used for financial claims. The hospital information systems contained patient-specific information regarding in-hospital stay, outpatient visits, daycare visits, intensive care, laboratory tests, radiology and administration of blood products. Medication use was derived from electronic patient charts for a random selection of patients (10% of all patients). This selection included both younger and older patients.

Unit prices of laboratory tests, radiology and other hospital activities were derived from national tariffs defined by the Dutch Health Authority (68). Reference unit prices were used for outpatient visits (€148), daycare treatment (€224), in-hospital stay (€712), intensive care days (€2,211) and blood products (69,70). Unit prices of medications were derived from the Pharmaceutical Compass (Z-index 2010). Unit prices of HLA-typing and donor search were obtained from Blommestein et al. (71). All unit prices included both capital and labor costs.

Costs were subcategorized into several cost groups: in-hospital stay, hospital visits, diagnostic procedures, medication, blood products, radiation, HLA-typing and donor search. Intensive care costs were included in the costs of in-hospital stay. Hospital visits consisted of daycare treatment, outpatient visits, emergency unit visits and other consults. Diagnostic procedures consisted of laboratory tests, radiology and other activities.

Follow-up data were not always available for one year, because only patients diagnosed in 2008 or 2009 were included in this study in order to calculate current treatment costs. If follow-up data were available for at least 100 days in patients alive without relapse, the costs were extrapolated to one year based on the average costs per day.

Missing values were imputed according to the average costs per in-hospital day during chemotherapy and transplantation. During follow-up, missing values were imputed according to the average costs per day spent in the hospital, including outpatient visits and daycare treatment. All costs were based on Euro 2010 cost data. Where necessary, costs were updated to 2010 according to the national consumer price index (72).

Statistical analysis

Mann-Whitney tests were used to test for significant differences in costs, in-hospital stay and treatment duration between treatment protocols, induction courses and consolidation treatments. A probability level <0.05 was considered significant. All analyses were performed using Microsoft Excel 2003 and SPSS 17.0.

RESULTS

Patients

In total, 202 patients were included in this study, of which 145 were treated according to the younger age protocol (Table 2.1). A second induction course was given to 127 (88%) younger and 40 (70%) older adults. Consolidation treatment was given to 126 patients. A few patients (N=8) received an allogeneic HSCT or UCB transplantation after only one induction course. Most patients received high-dose chemotherapy (N=47) as consolidation treatment. Sufficient follow-up data were available for 101 (70%) younger and 27 (47%) older adults.

Induction treatment

The average costs of the two induction courses were €46,807 for the first course and €42,395 for the second course (Table 2.2). Although the first induction course was significantly shorter than the second course, the total costs of the first course were significantly higher. This difference was mainly related to higher costs of diagnostic procedures as costs of diagnosis were included in the first course. In addition, blood products and medication costs were significantly higher during the first course. Although the chemotherapy dose was higher in the second course, the chemotherapy costs were lower due to the lower unit price of the anthracycline used in the second course. Costs of hospital visits were significantly lower in the first course. In-hospital stay did not differ

Table 2.1 Patients included in the different treatment phases

	All patients	"Younger age" protocol	"Older age" protocol
Induction course 1	202	145	57
Induction course 2	167	127	40
Consolidation treatment	126	126	-
High dose chemotherapy	47	47	-
Autologous SCT	18	18	-
Allogeneic SCT from sibling	35	35	-
Allogeneic SCT from MUD	21	21	-
Cord blood transplantation	5	5	-
Follow-up (1 year)	128	101	27
after induction treatment	27	-	27
after high dose chemotherapy	31	31	-
after autologous HSCT	16	16	-
after allogeneic HSCT from sibling	31	31	-
after allogeneic HSCT from MUD	19	19	-
after UCB transplantation	4	4	-

HSCT = hematopoietic stem cell transplantation
MUD = matched unrelated donor

Table 2.2 Costs of induction treatment (in 2010 Euros, 1 Euro = $1.3257)

	All patients		"Older" patients		"Younger" patients	
	Course 1 (N=202)	Course 2 (N=167)	Course 1 (N=57)	Course 2 (N=40)	Course 1 (N=145)	Course 2 (N=127)
Average time (days)	44*	60	48*	85†	42*	52
Mean costs						
(SE)						
In-hospital stay	24,333	24,837	24,551	25,622	24,247	24,589
	(575)	*(729)*	*(1,210)*	*(1,231)*	*(647)*	*(878)*
Hospital visits	342*	586	443*†	1,160*†	302	405
	(29)	*(58)*	*(66)*	*(158)*	*(30)*	*(49)*
Diagnostic activities	8,443*	5,583	8,704*	6,013†	8,341*	5,448
	(227)	*(247)*	*(408)*	*(370)*	*(273)*	*(303)*
Medication						
Chemotherapy	1,520*	1,354	366*†	766*†	1,974*	1,539
	(250)	*(138)*	*(32)*	*(40)*	*(67)*	*(95)*
Other medication	5,273*	4,006	6,873*†	4,982*†	4,645*	3,698
	(176)	*(189)*	*(433)*	*(380)*	*(149)*	*(218)*
Blood products	6,895*	6,029	6,547	7,067	7,032*	5,702
	(285)	*(341)*	*(445)*	*(764)*	*(356)*	*(378)*
Total costs						
Mean	**46,807***	**42,395**	**47,483**	**45,610†**	**46,541***	**41,382**
(SE)	***(1,076)***	***(1,362)***	***(2,113)***	***(2,463)***	***(1,251)***	***(1,609)***
Median	43,355	36,827	42,749	43,638	43,411	35,647

SE = standard error
* Significantly different from the costs in the second induction course ($p<0.05$)
† Significantly different from the induction costs of "younger" patients ($p<0.05$)

significantly between the two courses. On average patients were hospitalized 33.7 days during the first induction course and 34.1 days during the second course.

Not all patients received both induction courses. The majority of the patients who did not receive the second course was worse off as they died during the first course or were too ill to continue intensive treatment. Although patients receiving only one induction course were worse off, no significant difference in the total costs of the first induction course were found between patients receiving a second course or not (data not shown).

The total costs of the second induction course were significantly higher in older patients compared to younger patients. However, the duration of the second course was significantly longer in older patients; in fact, the total costs per treatment day did not differ significantly between the two age groups. No significant differences in total costs of the first induction course were found between the two age groups. In both induction courses, the chemotherapy costs were significantly higher in younger patients, while the costs of other medications were significantly lower. Duration of hospitalization did not differ significantly between the two age groups. The lower chemotherapy dose for older patients, and the subsequent decrease in complications, might be an explanation for this finding.

Consolidation treatment

The average costs of consolidation treatment were €34,225 for high-dose chemotherapy, €33,277 for autologous HSCT, €44,070 for allogeneic HSCT from sibling donor and €82,041 for allogeneic HSCT from MUD (Table 2.3). The costs of UCB transplantation were not calculated, as only five patients received this type of transplantation. No significant difference in total costs was found between high-dose chemotherapy, autologous HSCT and allogeneic HSCT from a sibling donor. However, some cost components differed significantly between the treatments. The costs of diagnostic procedures were significantly higher during autologous HSCT and allogeneic HSCT from a sibling donor compared to high-dose chemotherapy. The medication costs during autologous HSCT were significantly lower than the medication costs during high-dose chemotherapy. Although allogeneic HSCT from a sibling donor led to additional costs of radiation and HLA-typing, the total costs did not significantly differ from the costs of high-dose chemotherapy and autologous HSCT due to a significantly shorter hospital stay (22.5 days compared to 29.6 and 27.8 days, respectively) and lower blood product costs. The significantly shorter in-hospital stay during allogeneic HSCT from a sibling donor is likely due to outpatient strategies in allogeneic transplantations in modern times (73). An allogeneic HSCT from MUD was significantly more expensive than the other consolidation treatments. The higher costs were not only related to the costs of radiation, donor search

Table 2.3 Costs of consolidation treatment (in 2010 Euros, 1 Euro = $1.3257)

	High-dose chemotherapy	Autologous HSCT	Allogeneic HSCT sibling	Allogeneic HSCT MUD
	(N=47)	(N=18)	(N=35)	(N=21)
Average time (days)	100	106	101	128*‡
Mean costs				
(SE)				
In-hospital stay	21,247	19,944	17,007*†	18,682
	(914)	*(1,295)*	*(2,999)*	*(2,258)*
Hospital visits	1,477	1,706	2,038*†	2,767*†‡
	(148)	*(237)*	*(127)*	*(172)*
Diagnostic procedures	4,263	5,530*	8,099*	10,670*†‡
	(326)	*(512)*	*(1,213)*	*(778)*
Medication	3,389	3,050*	3,883	5,427*†‡
	(103)	*(162)*	*(466)*	*(566)*
Blood products	3,848	3,046	1,436*†	2,462*‡
	(346)	*(528)*	*(437)*	*(559)*
Radiation	-	-	1,638	1,610
			(100)	*(140)*
HLA-typing	-	-	9,968	9,968
			(NA)	*(NA)*
Donor search	-	-	-	30,456
				(NA)
Total costs				
Mean	**34,225**	**33,277**	**44,070**	**82,041*†‡**
(SE)	***(1,366)***	***(2,465)***	***(4,765)***	***(3,453)***
Median	33,031	31,951	37,394	83,165

HSCT = hematopoietic stem cell transplantation
MUD = matched unrelated donor
SE = standard error
NA = not available (mean costs derived from Blommestein et al.71)
***** Significantly different from high-dose chemotherapy ($p<0.05$)
† Significantly different from autologous HSCT ($p<0.05$)
‡ Significantly different from allogeneic HSCT from sibling donor ($p<0.05$)

and HLA-typing. The costs of hospital visits, diagnostic procedures and medication were also significantly higher during an allogeneic HSCT from MUD. In addition, the treatment duration was significantly longer during allogeneic HSCT from MUD compared to high-dose chemotherapy and allogeneic HSCT from a sibling donor.

Follow-up

The follow-up costs differed according to the preceding treatment (Table 2.4). The average 1-year follow-up costs were €11,740 after induction treatment, €5,856 after high-dose chemotherapy, €5,889 after autologous HSCT, €22,008 after allogeneic HSCT from sibling donor and €40,468 after allogeneic HSCT from MUD. The follow-up costs after allogeneic HSCTs (both from sibling donor and MUD) were significantly higher than the other follow-up costs, mainly because of higher costs of in-hospital stay, diagnostic procedures and hospital visits. The average days of hospitalization during follow-up ranged from 0 days after an autologous HSCT to 16.6 days after an allogeneic HSCT from MUD.

The follow-up costs included both patients who died or relapsed during the first year and patients alive without relapse. High follow-up costs might be caused by the high costs of patients who died during the first year, because it has been shown that the last year before death is the most expensive period in one's lifetime (74,75). In our study,

Table 2.4 Costs of follow-up (in 2010 Euros, 1 Euro = $1.3257)

	"Younger" patients				"Older" patients
Preceding treatment	High-dose chemotherapy	Autologous HSCT	Allogeneic HSCT sibling	Allogeneic HSCT MUD	Induction treatment
	(N=31)	(N=16)	(N=31)	(N=19)	(N=27)
Mean costs					
(SE)					
In-hospital stay	390	0	7,227*†§	16,687*†§	2,098
	(240)	*(0)*	*(2,556)*	*(5,455)*	*(961)*
Hospital visits	1,794	2,595	3,027*†	4,112*†§	2,702
	(234)	*(1,030)*	*(358)*	*(495)*	*(491)*
Diagnostic procedures	3,183	2,330	8,105*†§	14,395*†‡§	3,437
	(871)	*(396)*	*(1,119)*	*(1,918)*	*(756)*
Medication	200	164*	1,651*†§	2,450*†§	357
	(26)	*(58)*	*(232)*	*(416)*	*(87)*
Blood products	287§	800	1,998*†	2,822*†	3,146
	(184)	*(699)*	*(678)*	*(1,454)*	*(1,476)*
Total costs					
Mean	**5,856**	**5,889**	**22,008*†§**	**40,468*†‡§**	**11,740**
(SE)	***(1,172)***	***(2,000)***	***(4,345)***	***(8,299)***	***(3,630)***
Median	4,476	3,819	14,287	29,416	4,214

HSCT = hematopoietic stem cell transplantation
MUD = matched unrelated donor
SE = standard error
* Significantly different from follow-up after high-dose chemotherapy ($p<0.05$)
† Significantly different from follow-up after autologous HSCT ($p<0.05$)
‡ Significantly different from follow-up after allogeneic HSCT from sibling donor ($p<0.05$)
§ Significantly different from follow-up after induction treatment ($p<0.05$)

the mortality rate was significantly higher after induction treatment or allogeneic HSCT compared to the other treatments. Nevertheless, significant differences in follow-up costs were still apparent when the analysis was restricted to surviving patients. The higher follow-up costs after allogeneic transplantations are probably related to the treatment of graft-versus-host disease and infections. The average number of hospital visits after allogeneic transplantations were lower than expected according to treatment protocol, because patients who died during follow-up had significantly fewer hospital visits. These patients were more frequently admitted for in-hospital stay.

Composition of the costs of treatment

In-hospital stay accounted for 52-64% of the total costs of chemotherapy and autologous HSCT and 24-41% of the total costs of allogeneic transplantations. Other large cost components of allogeneic transplantations were HLA-typing and donor search (Figure 2.1A-B). The composition of the total follow-up costs differed according to the preceding treatment. In general, a large part of the follow-up costs consisted of diagnostic procedures (30-50%). The other follow-up costs were mainly related to in-hospital stay after an allogeneic HSCT and to costs of hospital visits after induction treatment, high-dose chemotherapy and autologous HSCT (Figure 2.1C).

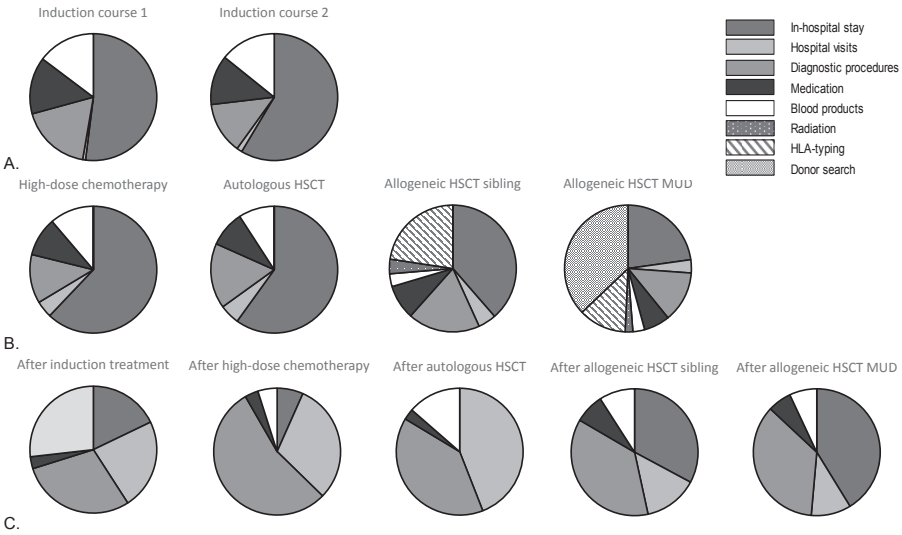


Figure 2.1 Composition of the total costs per treatment phase. Costs were subcategorized in the following subgroups: in-hospital stay, hospital visits, diagnostic activities, medication, blood products, radiation, HLA-typing and donor search. (A) Induction treatment cost, (B) consolidation treatment costs and (C) follow-up costs. HSCT = hematopoietic stem cell transplantation, MUD = matched unrelated donor

DISCUSSION

This study provides insight into the current costs of the initial treatment of AML. This study shows that the treatment costs of AML are substantial. The average costs of the initial treatment of one AML patient are €117,495. Previously estimated initial treatment costs were $80,030 in 1990 for younger patients and $52,048 in 1998 for older patients (respectively €80,109 and €59,630 in 2010 Euros) (62,63). In a Swedish study, the treatment costs, including relapse, of all AML patients were 356,911 Swedish Krona in 1990 (€74,508 in 2010 Euros) (61). The treatment costs calculated in this study were 43-67% higher than the previously estimated treatment costs. The largest increase in costs was found for diagnostic procedures, which increased with 80-220%. This large increase can be explained by both the use of new diagnostic technologies like real-time polymerase chain reaction and an increase in use of standard diagnostic procedures.

Only a few recent studies reported costs of parts of AML treatment. A French study estimated the total induction costs on €43,037 in 2005 Euros (€46,463 in 2010 Euros) (76). These costs are half of the induction costs calculated in our study. However, the French study differs from ours with respect to the treatment protocol and methodology of cost calculation. In the French study, a second course was only administered to patients without a CR after the first course (28% of all patients, compared to 83% in our study). In addition, costs of outpatient visits and related diagnostic procedures were excluded in the French study. These costs accounted for about 2.6% of the total costs per induction course in our study. The costs of sibling transplantations in our study were comparable with the costs estimated by Cordonnier et al. (77), which were €64,600 after high-dose conditioning (HDC) and €60,000 after reduced-intensity conditioning (RIC) in 2001 Euros, respectively (€75,718 and €69,980 in 2010 Euros).

It was not feasible to distinguish between HDC and RIC for allogeneic HSCT in our study. In the literature, some disagreement exists regarding differences in treatment costs between these conditioning regimens. Cordonnier et al. (77) did not find significant differences in costs, while Saito et al. (78) found significant lower transplantation costs after RIC due to a shorter hospital stay. The shorter hospital stay might also be related to the stem cell source. Peripheral blood transplants were significantly more administered in patients receiving RIC (78). As it has been shown that the hospital stay is shorter for patients receiving peripheral blood transplants (63), we do not expect that the type of conditioning has a large impact on treatment costs.

This study was a retrospective study with resource use collected from registered data. Due to the unavailability of data regarding the start and end of treatment phases, the

total period after hospitalization was defined as preparation of the following treatment phase. It is not expected that this definition will have large impact on the calculated costs, because the costs between treatment phases were relatively low.

Relapse costs were excluded in our study because only a small proportion of patients experienced a relapse in the available follow-time. It is expected that inclusion of relapse costs would especially increase the costs after high-dose chemotherapy and autologous HSCT, because the risk of relapse is higher after these treatments compared to allogeneic HSCT (79).

In general, in-hospital stay was the major cost driver in the treatment of AML. The main reason for the long hospital stay is that patients remain hospitalized until blood count recovery to avoid infections and bleeding. Several studies have shown that outpatient management after treatment is safe and feasible for a selection of patients (80,81). Outpatient management would not only reduce costs, but might also improve the quality of life of patients. Studies including more patients are warranted to investigate whether the hospital stay can be reduced without a negative impact on survival and quality of life.

The main objective of this study was to calculate costs of initial treatment of AML irrespective of patient characteristics. In addition, the costs of different treatment protocols were compared. As patients were not randomized between treatments, we cannot guarantee the comparability of the patients receiving different treatments. However, we do not expect large difference in patient characteristics, since treatment choice is mainly determined by the availability of a suitable donor. Additional analyses support this assumption as no significant differences in age were found between the post-remission treatments.

The results of this study can support hospitals in their negotiations with health insurers to receive adequate reimbursement for the treatment costs. Some concerns might exist regarding the transferability of the calculated costs to other countries. Although the included patients were only treated in the Netherlands, these patients were treated according to international guidelines (26). Further studies should investigate whether the costs of the second induction course differs between patients who achieved a CR after the first cycle or not. In addition, unit prices might differ between countries. However, a comparison of the costs with other recent cost calculations showed comparable results (76,77).

The calculated costs are representative estimates of current treatment costs, because resource use was derived from a sufficient number of patients treated according to current treatment protocols in several hospitals. Our results show large cost differences between the post-remission treatment options for AML patients. These differences should also be considered in treatment guidelines. While the effectiveness of AML treatments is established in randomized controlled trials, the cost-effectiveness is not always determined. Combining our results to current and future effectiveness studies might help to determine whether benefits of allogeneic HSCT outweigh the higher treatment costs.

ACKNOWLEDGEMENTS

The authors would like to thank Arjan Bandel and Kafong Cheung, Erasmus Medical Center Rotterdam; Koen Meijssen, VU University Medical Center; Ger Boonen and Evelien van Dijk, University Medical Center St. Radboud Nijmegen for providing the data from the hospital information systems.

Chapter 3

Mapping QLQ-C30, HAQ and MSIS-29 on EQ-5D

With Matthijs M. Versteegh, Jolanda J Luime, Mike Boggild, Carin A. Uyl-de Groot and Elly A. Stolk

Published in Medical Decision Making 2012(32): 554-568

ABSTRACT

Responses on condition-specific instruments can be mapped on EQ-5D to estimate utility values for economic evaluation. Mapping functions differ in predictive quality and not all condition-specific measures are suitable for estimating EQ-5D utilities. We mapped QLQ-C30, HAQ and MSIS-29 on the EQ-5D and compared the quality of the mapping functions with statistical and clinical indicators.

We used four datasets that included both the EQ-5D and a condition-specific measure to develop ordinary least squares regression equations. For the QLQ-C30, we used a multiple myeloma dataset and a non-Hodgkin's lymphoma one. An early arthritis cohort was used for the HAQ, and a cohort of patients with relapsing remitting or secondary progressive multiple sclerosis for the MSIS-29. We assessed the predictive quality of the mapping functions with the root mean square error (RMSE) and mean absolute error (MAE) and the ability to discriminate among relevant clinical subgroups. Pearson correlations between the condition-specific measures and items of the EQ-5D were used to determine if there is a relationship between the quality of the mapping functions and the amount of correlated content between the used measures.

QLQ-C30 had the highest correlation with EQ-5D items. Average %RMSE was best for QLQ-C30 with 10.9%, 12.2% for HAQ and 13.6% for MSIS-29. The mappings predicted mean EQ-5D utilities without significant differences with observed utilities and discriminated between relevant clinical groups, except for the HAQ model.

The preferred mapping functions in this study seem suitable for estimating EQ-5D utilities for economic evaluation. However, this research shows that lower correlations between instruments leads to less predictive quality. Using additional validation tests besides reporting statistical measures of error improves the assessment of predictive quality.

INTRODUCTION

Utility values (82), which are required to conduct cost-utility analyses, are usually measured by preference-based instruments like the EQ-5D, SF-6D or HUI III. Many clinical trials, however, use condition specific instruments, which do not incorporate preferences in the scoring algorithms, rather than preference-based instruments (83). Utility values can be estimated from the answers on a condition specific instrument when preference-based instruments are absent (32). This technique is called 'mapping' and primarily serves the purpose of 'rescuing' trial data for economic evaluation when a preference based instrument is absent. This paper presents a study that uses data from three condition specific questionnaires (Cancer: EORTC QLQ-C30 version 2; Arthritis: HAQ and Multiple Sclerosis: MSIS-29) to predict outcomes on a preference based instrument (the EQ-5D). We compare the quality of the mapping functions with statistical and clinical indicators and explore the influence of overlap in dimensions. Condition specific measures do not necessarily measure the same dimensions of health as a preference-based measure. The amount of overlap in dimensions between instruments is considered to be of influence on the predictive ability of the mapping function (84).

Mapping comes down to giving weight to different independent variables (items of a scale, the 'starting' measure) to predict the dependent variable (the 'target measure', e.g. utility index) through regression techniques. The independent variables may be sum scores, item scores, demographic variables and/or other (clinical) predictors of health. The purpose of such a mapping effort is to enable the estimation of a utility index in other datasets that do not have the target measure but do have the starting measure. The utility index, derived from the mapping effort, can be used to calculate the quality adjustment necessary for the computation of a quality adjusted life year (QALY). Previous studies have found that mapping is a feasible approach, but mapping models differ in predictive ability (32). Disadvantages of mapping are that the estimated utility indices have far greater errors for severe health states, and EQ-5D models tend to overpredict low utilities and underpredict high utilities (32,85). The performance of a mapping function may be tested by applying the algorithm to other (subset) data, which, like the sample on which the statistical association between dependent and independent variables was estimated, has data for both instruments.

Over the past few years several mapping algorithms have become available. In some circumstances, mapping may be the only way to get utility data, but the current growth in use of this strategy requires a more critical attitude towards the problems and promises of its application, as the quality of the mapping algorithms is highly variable (32). Assessing the quality of a mapping function, however, is not straightforward as different

indicators may hide certain flaws in a mapping function. First, an accurate prediction of mean utility values may cover that prediction errors may be larger for particular subgroups of patients. Second, (statistical) error indicators may not be easily interpretable without a comparator. Third, a measure of error does not directly reflect the external validity of a mapping function.

Another issue that deserves attention is the amount of overlap between dimensions of health covered by the descriptive systems of the starting and target measure. If the instruments focus on different dimensions of health (e.g. pain and mobility) they have little overlap and hence a low correlation between the items of the two measures, which may negatively influence the quality of the derived mapping function. The EQ-5D index values are the result of an algorithm that transforms the answers on five dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. To predict the index value, the variation of responses on those categories has to be predicted. A scale that only measures pain may have difficulty explaining variation in answers on a self-care domain. Consequently, the amount of overlap between instruments may influence the predictive ability of a mapping function.

This paper aims to present mapping functions for three condition-specific questionnaires onto EQ-5D suitable for use in economic evaluations, and to assess their quality through both statistical error measures and clinical indicators. We also test the hypothesis that the overlap of health domains is an important predictor for the predictive quality of a mapping function. If the amount of overlap between two instruments can be assessed at face value, it may inform a quick judgment about the expected quality of the derived mapping function.

METHODS

Instruments

Both the condition specific measures and the preference-based generic measure are patient reported outcome measures to assess health status. The measures have different properties as outlined below.

EQ-5D

The EQ-5D is a preference-based generic measure. It measures health related quality of life on five dimensions (mobility; self-care; usual activities; pain/discomfort and anxiety/depression) with three severity levels each. The measure was developed to provide a "simple abstracting device for use alongside other more detailed measures of health-

related quality of life to serve as a basis for comparing health outcomes" (86). The main outcome of interest is the derived utility (or preference) index; a single metric for quality of life derived by transforming the dimension scores with country-specific tariff. Utility values used in this study were computed using the Dutch EQ-5D tariff (87) and the UK EQ-5D tariff (88).

QLQ-C30

The EORTC-QLQ-C30 (version 2) is a cancer specific questionnaire and consists of 30 items, divided into three categories: functional scales (physical, role, emotional, cognitive and social functioning, for a total of 15 items), symptom scales (fatigue, nausea/vomiting, pain, dyspnea, sleep, appetite, constipation, diarrhea and financial difficulties, for a total of 13 items) and a global health status scale (two items). Scale sum scores are transformed so that a high score on the functional scales represents a high level of functioning, a high score on the symptom scales represents a high level of symptomatology and a high score on the global health status represents a high quality of life (89). At face value it seems all health domains of EQ-5D are present in the QLQ-C30. The QLQ-C30 has been successfully mapped on the EQ-5D (UK tariff) before (90), but not for Dutch utilities nor for a lymphoma population.

HAQ

The Health Assessment Questionnaire (HAQ) is a widely used questionnaire in the field of rheumatology. The HAQ assesses the functional ability of patients using 20 items across eight domains (dressing, rising, eating, walking, hygiene, reach, grip and usual activities) (91). Items are scored on a 4-level disability scale from zero to three, where three represents the highest degree of disability. Scores are adjusted for the use of aids or devices and averaged into dimension sum scores and an overall disability index value, which represents the extent of functional ability of the patient. Values between one and two represent moderate to severe disability (91). A face value judgment of the items and sum scores of HAQ suggests the EQ-5D dimensions pain/discomfort and anxiety/depression are not represented in the HAQ. The HAQ has been mapped on the EQ-5D before and was able to predict mean utility values (92), but such a function has not been estimated for Dutch utilities.

MSIS-29

The MSIS-29 is a Multiple Sclerosis (MS) specific questionnaire with 29 items developed through reducing an item pool of 129 items concerning the health impact of multiple sclerosis (93). The MSIS-29 is a self-reported measure which measures the physical and psychological impact of MS on individuals. Items measure disease impact due to limitations in the past two weeks, scored on five levels from 'not at all' to 'extremely'. The first

20 items (physical scale) and the last 9 items (psychological scale) form two summary scores transformed to a 0-100 scale. The MSIS-29 has not been previously mapped to our knowledge. The EQ-5D dimensions mobility, self-care, and usual activities are not explicitly represented as dimensions of the scale, but the dimensions are tapped into by items like 'difficulty moving about indoors', 'having to depend on others to do things for you' and 'limitations in your social and leisure activities at home'. MSIS-29 items and sum scores suggests the EQ-5D dimension 'pain/discomfort' is not present in the MSIS-29.

Table 3.1 Patient characteristics

Sample		Development set	Test set 1	Test set 2
QLQ-C30	N	723 (pooled)	789 (pooled)	
	Age (range)	54 (37 - 64)	72 (65-84)*	
EQ-5D†	Mobility % 1/2/3	56.7 / 41.4 / 1.9	48 / 47.3 / 4.7	
	Self-care % 1/2/3	85.8 / 12.8 / 1.4	81.4 / 13.9 / 4.7	
	Usual activities % 1/2/3	30.1 / 51.1 / 18.8	38.1 / 43.3 / 18.6	
	Pain/Discomfort % 1/2/3	39.6 / 59.0 / 1.4	52.2 / 42.9 / 4.9	
	Depression/Anxiety % 1/2/3	69.4 / 29.6 / 1.0	70 / 29 / 1.0	
	EQ-5D-index	.742 (.21)	.735 (.26)	
	Cancer type	Multiple Myeloma	Non-Hodgkin lymphoma	
QLQ-C30	Physical functioning	64 (24.6)	57.3 (26.8)*	
(pooled)	Role functioning	59.5 (28.9)	57.4 (31.5)	
	Emotional functioning	82.8 (18.9)	81.3 (20.6)	
	Cognitive functioning	82 (20.8)	81.9 (23.7)	
	Social functioning	76.2 (25.8)	75.7 (28.6)	
	Global health	68.7 (18.0)	62 (21.7)*	
	Fatigue	35.7 (25.0)	44.7 (29.4)*	
	Nausea / Vomiting	6.1 (14.3)	8 (16.9)*	
	Pain	25.2 (24.7)	18.7 (26.2)*	
	Dyspnoea	16.1 (24.9)	24.8 (28.9)*	
	Sleep	21.1 (27.3)	28.7 (31.8)*	
	Appetite	16 (27.2)	21.9 (32.6)*	
	Constipation	4 (15.4)	11.8 (22.8)*	
	Diarrhea	8.3 (18.7)	7 (18.5)*	
	Financial difficulties	12.5 (23.0)	6.3 (16.9)*	
HAQ	N	186	132	
	Age (range)	51 (16 - 82)	55 (25 – 78)	
EQ-5D†	Mobility % 1/2/3	38.2 / 61.8 / 0.0	44.4 / 55.3 / 0.0	
	Self-care % 1/2/3	68.5 / 30.4 / 1.1	58.0 / 41.2 / 0.8	
	Usual activities % 1/2/3	33.9 / 60.1 / 6.0	27.5 / 67.9 / 4.6	

Table 3.1 Patient characteristics (continued)

Sample		Development set	Test set 1	Test set 2
	Pain/Discomfort % 1/2/3	8.1 / 78.4 / 13.5	3.8 / 77.3/ 18.9	
	Depression/Anxiety % 1/2/3	72.4 / 25.4 / 2.2	80.3 / 17.4 / 2.3	
	EQ-5D-index	0.67 (0.24)	0.64 (0.26)	
	DAS 28	4.34 (1.30)	4.30 (1.27)	
	HAQ-DI	0.75 (0.65)	0.81 (0.65)	
HAQ	Dressing & Grooming	0.64 (0.72)	0.76 (0.75)	
	Arising	0.71 (0.77)	0.72 (0.75)	
	Eating	0.84 (0.84)	0.95 (0.83)	
	Walking	0.64 (.85)	0.58 (0.81)	
	Hygiene	0.70 (0.85)	0.79 (0.86)	
	Reach	0.68 (0.80)	0.74 (0.81)	
	Grip	0.87 (0.86)	0.93 (0.86)	
	Activities	0.93 (0.86)	1.04 (0.88)	
MSIS-29	N	661	339	295
	Age (range)	40 (18-88)	40 (18 -87)	41 (18-88)
EQ-5D†	Mobility % 1/2/3	21.2 / 76.6 / 2.2	26.7 / 70.9 / 2.4	25.0 / 74.3 / 0.7
	Self-care % 1/2/3	62.4 / 35.8 / 1.8	68.2 / 29.4 / 2.4	63.0 / 35.6 / 1.4
	Usual activities % 1/2/3	21.1 / 70.9 / 8.0	22.9 / 69.4 / 8.3	23.5 / 71.3 / 5.2
	Pain/Discomfort % 1/2/3	25.3 / 67.0 / 7.7	27.5 / 61.5 / 11.0	24.7 / 68.2 / 7.1
	Depression/Anxiety % 1/2/3	40.7 / 52.1 / 7.2	41.1 / 50.9 / 8.0	39.2 / 56.3 / 4.5
	EQ-5D-index UK / NL	0.58 (0.29) / 0.63 (0.62)	0.57 (0.31) / 0.61 (0.20)	0.60 (0.25) / 0.65 (0.23)
	Type of MS‡	RR = 81% SP = 19%	RR = 81% SP = 19%	RR = 82% SP = 18%
MSIS-29	Physical scale	47.1 (25.7)	45.4 (25.8)	44.9 (25.7)
	Psychological scale	45.4 (25.6)	44.7 (24.9)	44.4 (25.6)

* Significant difference (2-tailed t-test $p < .05$) with development set
† % at level 1 / 2 / 3 (not at all / some problems / extreme)
‡ RR= Relapsing Remitting, SP = Secondary Progressive

Population

The EQ-5D and condition specific measure data were retrieved from three different datasets described below. An overview of characteristics of the populations is presented in Table 3.1.

QLQ-C30: the HOVON study

Data for QLQ-C30 model were taken from two separate studies carried out by the Dutch association for hematology/oncology in adults (HOVON). The HOVON 24 (94)

and HOVON 25 (95) studies are randomized clinical trials that measure the effectiveness of different treatments in respectively patients with previously untreated multiple myeloma (HOVON 24) and previously untreated non-Hodgkin lymphoma patients (Ann Arbor stages II to IV, or intermediate or high-grade malignancy). The sample size of the clinical trial is larger than the sample size of patients that had concluded both an EORTC QLQ-C30 and an EQ5D instrument.

The mapping algorithm was developed on the multiple myeloma sample, and tested on the Non-Hodgkin lymphoma sample. The database for the multiple myeloma sample contained 137 patients at baseline with 6 follow-up measurements (three early follow-ups were missing), the latest two years after baseline. To increase the number of data-points per EQ-5D utility value – samples are often short on severe health states and thus lack predictive ability in the lower range – follow up measurements were pooled for the development of prediction models. The database for the Non-Hodgkin's sample contained 108 patients at baseline and had seven time points at which the EORTC QLQ C-30 was administered. Three time-points were after the second, fourth and sixth treatment cycle, and four follow-up measurements were at baseline, three, six, ten and eighteen months after baseline. Predictive ability of the mapping is assessed per time point.

HAQ: the REACH study

The HAQ data were taken from the Rotterdam Early Arthritis CoHort (REACH) with patients recruited from the Erasmus Medical Centre in the Netherlands, with arthritis. As can be inferred from the name, one of the aims of the study is early detection of rheumatoid arthritis. Data are collected through three outpatient visits, during which respondents filled out a booklet of self-reported questionnaires including HAQ, Hospital Anxiety and Depression Scale (HADS), Short Form-36 (SF-36) and EQ-5D. The mapping algorithm is developed on a randomly drawn sub-sample of the dataset (n=186) and tested on a remaining sample (n=132) for which most of the data of the questionnaires were available, both at baseline. A randomly drawn subsample is not expected to deviate much from the remaining sample.

MSIS-29: the Multiple Sclerosis Risk Sharing Scheme Monitoring Study

The MSIS-29 data were taken from the Multiple Sclerosis Risk Sharing Scheme Monitoring Study in the UK (96), which aims to analyze the long term cost effectiveness of disease modifying treatments in patients with MS. Data for the study were collected from MS patients in 70 specialist MS centers in the UK and includes both relapsing remitting and secondary progressive MS patients. Cross-sectional cost and utility data for a subset of these MS patients was collected to enhance economic analysis. The MSIS-29 was administered once (n= 1,295); hence there are no different time-points for this mea-

sure. The mapping algorithm was developed on a randomly drawn sub sample (n=661) and tested on two samples randomly drawn from the remaining respondents (n=339 and n=295). As with the HAQ, the randomly drawn samples are not expected to deviate much from the rest of the data.

Analysis

Pearson's r and Spearman's rho determined the amount of overlap between the instruments. Ordinary least squares regression models will be fitted to the data to generate mapping algorithms. All models were developed for the Dutch value set of the EQ-5D (87). As the MS dataset is based on the English versions of the MSIS-29 and EQ-5D, the algorithm was also estimated for the UK value set (88).

The models in this paper serve the purpose of calculating mean utility scores applicable in economic evaluation. The mapping models generate individual utility values, but these have uncertain estimates (97). Even if the estimates were less uncertain, the EQ-5D is a tool for economic evaluation and priority setting in health care (98), not a key health indicator of individual health. Thus only aggregated utilities derived from mapping studies are to be used. Analyses were run in STATA 10.0 and SPSS 17.0.

Model development

This study follows the suggestions of a recent review of mapping studies which suggests estimating different types of models with increasing complexity and decreasing assumptions about the properties of the data (32). The strategy implies that a range of models is estimated, with increasing levels of complexity. All models that we developed aimed to compute EQ-5D utility scores from the condition specific measures with ordinary least squares regression. An alternative approach would be to predict the dimension scores of EQ-5D through multinomial regression. This implies that the models predict occurrence of level 1, 2 and 3 responses on the EQ-5D from which one indirectly computes the utility scores. We chose ordinary least squares regression to predict the EQ-5D utility index as predicting dimension levels has been shown to have equal or worse predictive performance (32) for both QLQ-C30 (90) and HAQ (92).

All models were developed in similar order, relaxing the assumptions about the properties of the condition specific measures in each step, while eyeballing model performance (described below) as guidance for model selection. First the EQ-5D utility index was regressed on sum scores, then on item scores treated as continuous variables. Treating items as continuous variables assumes interval properties of the questionnaire, because only one coefficient is calculated for all changes in responses of one item (i.e. On the question 'have you been bothered by problems sleeping?' moving from answer

category 'not at all' to 'a bit' receives the same decrement as moving from adjacent categories 'quite a bit' to 'extremely'). Lastly, by treating item scores as categorical (dummy) variables, this assumption was relaxed. Dummies are computed to let each subsequent level represent worse health compared to the reference category 'no problems'. Models were developed using backward selection procedures with probability of F to remove a variable at 0.10.

Models were required to be logically consistent meaning that a worse score on an item should lead to a larger utility decrement. For reasons of parsimony models were reduced to the smallest models that have similar predictive performance as larger models. In the final models achieving parsimony was attempted by merging item categories with the logically adjacent answer category when they a) did not meet probability <0.10 of F to remove or b) were logically inconsistent.

Clinical variables or additional questionnaires may correlate with a given dimension (muscle strength may capture mobility). Such variables may improve the mapping but might not be available as data in all clinical trials, potentially limiting the use of a function. As the HAQ is a measure of disability and not a quality of life questionnaire we expected that HAQ based models might not yield favorable prediction results. We therefore selected five variables in our dataset that theoretically would correlate with EQ-5D dimensions. The variables are: sum scores of the HADS, sum scores of the SF-36 and the Disease Activity Score (DAS28) and a count of tender and swollen joints. Analysis started with all predictor variables as regression coefficients and proceeded through backward step wise regression with probability of F to remove at 0.10. Missing values were not imputed in any of the datasets.

Model performance

There are no strict criteria for a utility function to be acceptable (32). QALYs are computed using mean scores, thus interesting indicators are differences between the predicted mean utility score and the observed mean utility score. However, correct prediction of a mean may still hide differences between individual observed and predicted values across the entire scale. Model performance is therefore reported as root mean squared error (RMSE) and mean absolute error (MAE) of the predicted EQ-5D utility values (32,99). RMSE is the root of the average of the squared differences between observed and predicted values, while MAE is the average of the roots of the squared difference between observed and predicted values. Lower values of RMSE and MAE indicate better model performance. As RMSE averages the squared differences it is sensitive to extreme deviations from the mean (e.g. outliers) and thus always equal to or larger than MAE. There is no definition of what level of RMSE or MAE is a threshold for model acceptance (92).

Besides that, RMSE and MAE are not comparable for models with different preference-based instruments as dependent variables or models with a different range of observed values since larger scale size usually leads to a larger error figure. For instance: UK EQ-5D index values have a measurement range of 1.59 compared to 1.33 for the Dutch values. Consequently RMSE is also reported as a percentage of the scale size, the normalized RMSE (100). It is expected that a higher prediction error is positively associated with less overlap between EQ-5D and the disease specific instrument.

Discriminant validity

Statistical measures such as RMSE and MAE may be difficult to interpret and can be small because of the method of testing, caused, for instance by using a randomly drawn subsample which does not deviate much from the development sample. Therefore, the validity of predicted index values is also inspected by checking the ability of the predicted values to discriminate between relevant clinical groups. Mean scores of observed and predicted EQ-5D index values are calculated per category of a relevant clinical indicator. For the cancer data this clinical indicator is the doctor reported World Health Organization performance status (or ECOG score) which distinguishes 6 categories from 0 (asymptomatic) to 6 (death). For the arthritis data the clinical indicator is the DAS28 and is based on a count of tender joints and the erythrocyte sedimentation rate (ESR). It can be used to distinguish between high, moderate and low disease activity and remission. Lastly, for multiple sclerosis the data are compared to the categories of the Expended Disability Status Scale (EDSS) which range from 0 (no neurological deficit) and 10 (death). Pearson r is used to measure the (linear) correlation between the clinical indicators and the observed and predicted scores.

Table 3.2 Pearson's correlation matrix between sum scores and EQ-5D dimensions

	EQ-5D domain				
QLQ-C30†	mobility	self care	usual activities	pain/ discomfort	anxiety/ depression
Physical functioning	-0.67**	-0.48**	-0.64**	-0.45**	-0.27**
Role functioning	-0.54**	-0.38**	-0.78**	-0.47**	-0.29**
Emotional functioning	-0.22**	-0.20*	-0.30**	-0.24**	-0.70**
Cognitive functioning	-0.20*	-0.17*	-0.28**	-0.27**	-0.37**
Social functioning	-0.49**	-0.34**	-0.55**	-0.30**	-0.40**
Global health status	-0.36**	-0.18*	-0.47**	-0.44**	-0.39**
Fatigue	0.35**	0.20*	0.51**	0.39**	0.34**
Nausea and vomiting	0.03	-0.08	0.11	0.21*	0.09
Pain	0.40**	0.17*	0.41**	0.76**	0.19*
Dyspnoea	0.11	0.13	0.26**	0.08	-0.02
Sleep	0.08	0.10	0.19*	0.19*	0.30**
Appetite loss	0.25**	0.15	0.28**	0.25**	0.27**
Constipation	0.05	0.04	0.05	0.26**	0.04
Diarrhoea	0.21*	0.07	0.14	0.08	0.10
Financial difficulties	0.14	0.22**	0.22**	-0.01	0.12
HAQ	mobility	self care	usual activities	pain/ discomfort	anxiety/ depression
Dressing	0.29**	0.63**	0.42**	0.41**	0.17**
Rising	0.44**	0.51**	0.45**	0.37**	0.20**
Eating	0.27**	0.52**	0.44**	0.36**	0.11*
Walking	0.52**	0.48**	0.39**	0.32**	0.21**
Hygiene	0.35**	0.62**	0.41**	0.39**	0.19**
Reach	0.35**	0.54**	0.39**	0.44**	0.23**
Grip	0.27**	0.45**	0.36**	0.40**	0.15**
Usual activities	0.42**	0.54**	0.51**	0.45**	0.18**
MSIS29	mobility	self care	usual activities	pain/ discomfort	anxiety/ depression
Physical scale	0.67**	0.59**	0.67**	0.50**	0.40**
Psychological	0.43**	0.37**	0.46**	0.45**	0.68**

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

† Time point = baseline

Shaded cell > 0.55

Table 3.3 Model parameters

	Predictors	*B*	*SE*	*p*		Predictors	*B*	*SE*	*p*		Predictors	*B*	*SE*	*p*
QLQ-C30	(Constant)	0.978	,0.008	0.000	**HAQ**	(Constant)	0.,527	0.113	0.000	**MSIS-29**	(Constant)	,0.956	0.021	0.000
Model 4	QLQ1	-0.030	0.010	0.002	**Model 3**	HAQ8	-0.038	0.024	0.106	**Model 4**	MSI3_5	-0.075	0.022	0.001
R^2 (A) =0.74	QLQ2	-0.025	,0.009	,0.007	R^2 (A) =0.54	HADS_D	-0.017	0.005	0.001	R^2 (A) =0.58	MSI5_2	-,0.046	0.021	0.030
R^2 (B) = 0.73	QLQ3	-0.045	0.010	,0.000	R^2 (B) = 0.52	SF36_PF	0.001	0.001	0.107	R^2 (B) = 0.49	MSI5_3	-0.055	0.025	0.029
RMSE =0.13	QLQ4	-0.069	,0.011	,0.000	RMSE=,0.15	SF36_RP	-,0.001	,0.000	0.032	RMSE =0.19	MSI5_4	-0.056	0.028	0.046
	QLQ5	-0.159	,0.016	0.000		SF36_BP	0.005	0.001	0.000		MSI5_5	-,0.129	0.035	0.000
	QLQ6_1	-0.037	,0.010	,0.000		SF36_RE	0.001	0.000	,0.024		MSI6_2	-0.062	0.025	0.013
	QLQ6_2	-0.077	,0.015	,0.000		DAS28	-,0.019	0.012	0.113		MSI6_3_4	-0.071	0.028	0.013
	QLQ6_3	-0.187	0.019	,0.000							MSI6_5	-0.134	0.036	0.000
	QLQ7_2_3	-0.020	0.011	,0.084							MSI10_2_3	-0.049	0.016	0.003
	QLQ9_1_2	-0.076	,0.007	0.000							MSI10_4	-0.084	0.023	0.000
	QLQ9_3	-0.267	0.019	,0.000							MSI10_5	-0.082	0.030	0.006
	QLQ23_1	-0.020	0.008	0.015							MSI15_2	-0.036	0.019	0.064
	QLQ23_2	-0.028	0.016	,0.070							MSI15_3	-0.068	0.023	0.004
	QLQ23_3	-0.267	0.,048	0.000							MSI15_4	-0.068	0.025	0.006
	QLQ24_1	-0.071	0.009	,0.000							MSI15_5	-,0.105	0.033	0.001
	QLQ24_2_3	-0.144	0.015	0.000							MSI21_3	-0.031	0.018	0.084
	QLQ27_2	-0.041	,0.010	,0.000							MSI21_4	-0.044	0.021	0.037
	QLQ27_3	-0.063	,0.016	0.000							MSI21_5	-0.141	,0.029	0.000
											MSI22_5	-0.098	0.024	0.000
											MSI28_4	-0.043	0.020	0.031
											MSI28_5	-0.042	0.026	,0.100
											MSI29_2	-0.051	0.018	0.006
											MSI29_3_4	-0.070	0.020	0.000
											MSI29_5	-0.204	0.029	0.000

R^2 (A) = Adjusted R^2 in development sample

R^2 (B) = Adjusted R^2 in test sample

RMSE = root mean square error/HADS_D= Depression sum score of HADS / SF36_PF=Physical Functioning / SF36_RP=Role-functioning / SF36BP=Bodily Pain/ SF36RE = Role-emotional / DAS28 = Disease Activity Score

Table 3.4 MSIS-29 UK model

	Predictors	*B*	*SE*	*p*
Model 1	(Constant)	0.949	0.024	0.000
R^2 (A) =0.57	msi3_3_4	-0.055	0.021	0.009
R^2 (B) = 0.49	msi3_5	-0.144	0.031	0.000
RMSE =0.19	msi5_5	-0.063	0.033	0.054
	msi6_2	-0.067	0.029	0.020
	msi6_3_4	-0.075	0.032	0.019
	msi6_5	-0.142	0.040	0.000
	msi10_2_3	-0.060	0.018	0.001
	msi10_4	-0.116	0.025	0.000
	msi10_5	-0.130	0.033	0.000
	msi15_5	-0.070	0.031	0.023
	msi16_3_4	-0.042	0.022	0.057
	msi16_5	-0.065	0.032	0.041
	msi18_2	-0.061	0.028	0.030
	msi18_3	-0.107	0.033	0.001
	msi18_4_5	-0.118	0.036	0.001
	msi21_5	-0.131	0.028	0.000
	msi26_3_4_5	-0.033	0.018	0.068
	msi29_3_4	-0.038	0.019	0.044
	msi29_5	-0.188	0.027	0.000

R^2 (A) = Adjusted R^2 in development sample
R^2 (B) = Adjusted R^2 in test sample

RESULTS

Pearson correlations indicate that the condition specific measures differ in the amount of overlap with EQ-5D dimensions (Table 3.2). Table 3.2 suggests that Pearson correlations with EQ-5D dimensions were higher for QLQ-C30 and MSIS-29 than for HAQ and were nearly identical to Spearman ρ. For instance, none of the HAQ dimensions had a correlation coefficient >0.23 with the EQ-5D dimensions 'anxiety / depression', whilst this dimensions correlates with -0.70 and 0.68 for respectively QLQ-C30 and MSIS-29. The QLQ-C30 has the highest correlations with EQ-5D dimensions. Based on these results, we would expect the mapping functions based on QLQ-C30 to outperform those based on the HAQ and MSIS-29.

Mappings

The best functioning models and their performance are summarized in Table 3.3 and Table 3.4 (for all developed models see Supplementary material) and further discussed

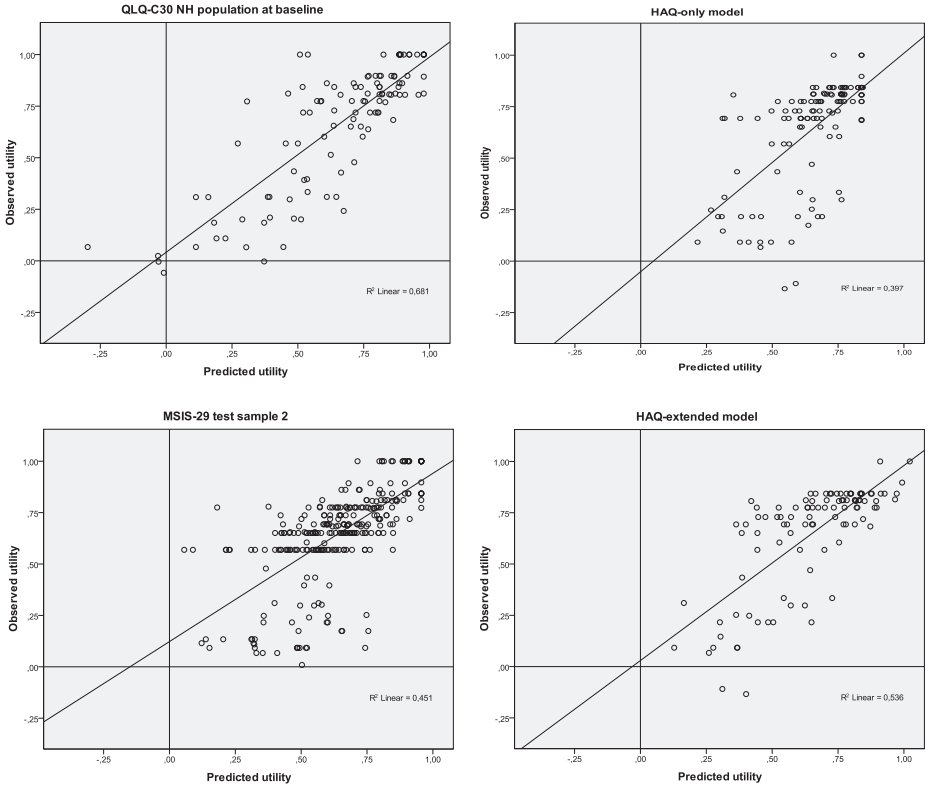


Figure 3.1 Scatter plots of observed and predicted values

below. QLQ-C30 model 4, HAQ model 3, and MSIS-29 model 4 meet the requirements of logical consistency, significance of predictors, parsimony and were able to predict mean utility values in the test samples (Table 3.5). Scatter plots of the predicted compared to observed values at baseline in the test samples (MSIS-29 test sample 2) are presented in Figure 3.1. Box plots of prediction errors per EQ-5D utility category are presented in Figure 3.2.

QLQ-C30

Predictions were better when the assumptions concerning data were relaxed. As expected from the relatively low correlation with EQ-5D dimensions, the sum score of cognitive functioning scale was excluded after backward selection in model 1, as cognitive functioning is not represented in the EQ-5D. Using items as continuous predictor variables (model 2) reduced prediction errors, but performed worse than model 3 which used items as dummy variables. In model 3, after backward selection, it was decided to remove items from the model based on illogical signs (item 10, 22 and 25), at the cost of 2% explained variance. Dummy model 3 consists of items 1 to 5 (physical functioning); 6

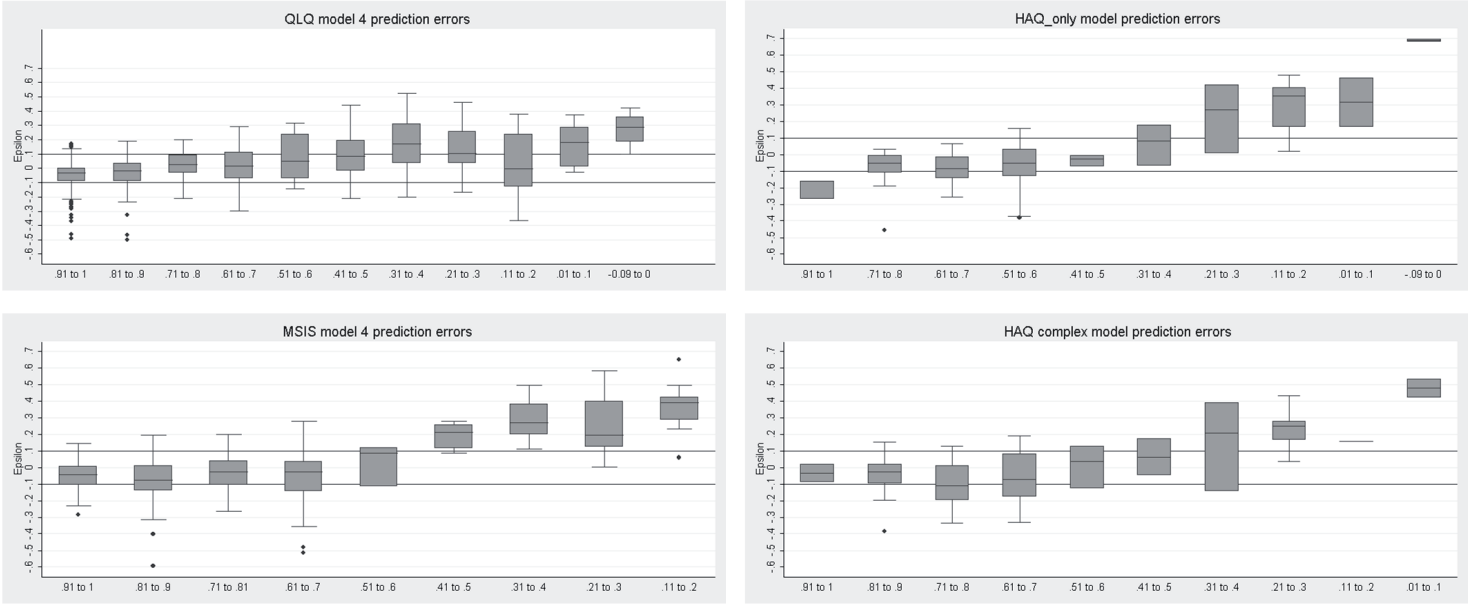


Figure3.2 Box plots of Epsilon (Observed – Predicted) per EQ-5D utility category

and 7 (role functioning); 9 (pain); 16 (constipation); 23 & 24 (emotional functioning) and item 27 (social functioning). To achieve a more parsimonious model, non-significant, and remaining significant but illogically ordered (items 7 and 9) dummy categories were merged and item 16 (constipation) was dropped altogether, without effecting the predictive performance of the model. The model developed in the multiple myeloma patient sample was able to predict utilities in the non-Hodgkin's sample. The largest RMSE value is at baseline and 10 months follow-up which is due to an outlier as can be seen from the relatively low MAE which is less sensitive to outliers. When the model could not predict a utility value because the non-Hodgking sample had missing values in QLQ-C30 responses, or when the utility value from the EQ-5D was missing, cases were excluded leading to minimal differences in RMSE (smaller than 0.01) compared to including missing data-points as well.

Table 3.5 Summary of model performance in test samples

Model	Observed mean EQ-5D (SD)	Predicted mean EQ-5D (SD)	RMSE (normalized for range)	MAE	R^2	Min - Max observed	Min - Max predicted
QLQ-C30							
Baseline	0.66 (0.30)	0.66 (0.26)	0.16 (12.0%)	0.12	0.75	-0.06 - 1	-0.30 - 0.98
2nd treatment cycle	0.70 (0.26)	0.71 (0.22)	0.13 (9.7%)	0.10	0.79	-0.13 - 1	0.10 - 0.98
4th treatment cycle	0.72 (0.25)	0.72 (0.21)	0.12 (9.0%)	0.08	0.79	-0.09 - 1	0.13 - 0.98
6th treatment cycle	0.70 (0.26)	0.69 (0.22)	0.15 (11.3%)	0.10	0.75	-0.13 - 1	-0.06 - 0.98
3 months follow-up	0.77 (0.26)	0.77 (0.21)	0.10 (7.5%)	0.07	0.82	-0.13 - 1	-0.03 - 0.98
6 months follow-up	0.80 (0.20)	0.79 (0.18)	0.11 (8.3%)	0.07	0.74	0.00 - 1	-0.03 - 0.98
10 months follow-up	0.77 (0.27)	0.80 (0.19)	0.16 (12.0%)	0.09	0.68	-0.33 - 1	0.02 - 0.98
18 months follow-up	0.81 (0.18)	0.82 (0.18)	0.09 (6.7%)	0.06	0.8	0.22 - 1	0.01 - 0.98
HAQ							
HAQ-only model	0.64 (0.26)	0.65 (0.16)	0.17 (12.2%)	0.07	0.39	-0.13 - 1	0.22 - 0.84
Extended model	0.64 (0.26)	0.65 (0.20)	0.15 (10.8%)	0.04	0.54	-0.13 - 1	0.13 - 1.02
MSIS-29							
Test sample 1	0.62 (0.28)	0.62 (0.27)	0.20 (14.4%)	0.16	0.49	-0.13 - 1	-0.05 - 0.96
Test sample 2	0.65 (0.23)	0.65 (0.19)	0.18 (12.9%)	0.13	0.49	0.01 - 1	0.06 - 0.96
MSIS-29 (UK)							
Test sample 1	0.57 (0.31)	0.59 (0.22)	0.22 (13.8%)	0.16	0.49	-0.32 - 1	-0.14 - 0.95
Test sample 2	0.60 (0.26)	0.60 (0.21)	0.18 (11.3%)	0.13	0.49	-0.17 - 1	0.00 - 0.95

HAQ

As the sample was focused on early arthritis the dummy items often lacked respondents which scored the lowest answer category (level 4). In a dummy model, the variables that represent the fourth category mostly score 0, yielding it impossible to estimate a decrement for a level 4 answer. Consequently a model with summed dimension scores had better RMSE scores in the test sample. Removing the insignificant sum variables in the prediction model did not improve predictions in the test sample. After backward stepwise regression, the model with the a priori selected predictor variables (extended model) performed better in terms of RMSE, MAE and R^2 and was capable of predicting a wider range of EQ-5D index values as can be seen in Figure 3.1 and Table 3.5. Stepwise selection of variables for the extended model resulted in the removal of all HAQ sum scores except usual activities. Tender joint count and swollen joint count did not contribute to the model but are represented indirectly through the DAS28. The other variables are the transformed sum scores of the SF-36 (physical functioning, role-physical, bodily pain, and role-emotional) and the depression sum score of the HADS. Added variables had significant Pearson correlations between 0.47 and 0.58 with at least one of EQ-5D dimensions. Adding a squared term for the SF-36 bodily pain score did not reduce prediction errors. The extended model failed to predict 27 of the 132 EQ-5D index values due to missing data for one or more of the prediction variables.

MSIS-29

For the Dutch value set, MSIS-29 models 2 to 4 with items as predictors performed better than sum score model 1. Treating items as dummy variables did not reduce prediction error in terms of RMSE and MAE. However, the continuous model seemed illogical with an intercept of 1.17 while a value of 1 represents full health in the QALY-model. Model 3 contained many insignificant and illogical items, but removing them increased prediction errors. However, a more parsimonious model with significant predictors could be developed without losing predictive performance through merging categories. Only MSIS-29 item 10 and 28 could not be further reduced into smaller categories without losing predictive performance. Consequently 4 variables in the model are not logically ordered, however the largest difference is 0.002 between MSI10_4 and MSI10_5.

Results were similar for the UK value set. The most parsimonious model is presented in Table 3.4, which had, however, slightly larger prediction errors. The final algorithm contained 10 items, 7 items from the physical dimension and 3 from the psychological dimension.

Table 3.6 Comparison of predicted and observed EQ-5D index values by clinical indicators

QLQ-C30

WHO	N (summed)	EQ-5D index	Mapped EQ-5D model 4	QLQ sum score global health	
0	356	0.75	0.75	63	
1	304	0.76	0.74	64	
2	96	0.69	0.72	59	
3	27	0.37	0.42	41	
Pearson's r		-0.19**	-0.19**	-0.095**′	

MSIS-29

EDSS	N	EQ-5D index	Mapped EQ-5D model 4	MSIS-PHY	MSIS-PSY
0	9	0.81	0.81	34	21
1	25	0.73	0.72	40	22
2	71	0.74	0.70	44	22
3	51	0.58	0.59	55	26
4	56	0.63	0.58	57	26
5	28	0.61	0.54	64	26
6	53	0.55	0.49	69	28
7	25	0.40	0.40	74	28
Pearson's r		-0.38**	-0.47**	0.57**	0.23**

HAQ

DAS28	N	N Model 3	EQ-5D index	Mapped EQ-5D model 1	Mapped EQ-5D model 3
Remission	11	9	0.76	0.76	0.83
Low DA	15	12	0.70	0.68	0.73
Moderate DA	70	59	0.67	0.68	0.67
High DA	27	23	0.51	0.54	0.49
Pearson's r			-0.37**	-0.52**	-0.57**

** *p<0.00*

Discriminant validity

The preferred models were tested for their ability to discriminate between relevant clinical subgroups (known-groups analysis). Results are satisfactory for all the preferred mapping models (QLQ-C30 model 4, HAQ model 3, MSIS-29 model 4 and the MSIS-29 UK model) as presented in Table 3.6. For the QLQ-C30 the predicted values follow a similar pattern to the observed EQ-5D values. Both the predicted and the observed EQ-5D values can hardly distinguish between WHO categories 0 (fully active) and 1 (cannot do heavy physical work but can do everything else), but neither can the self-assessed global health sum score from the QLQ-C30.

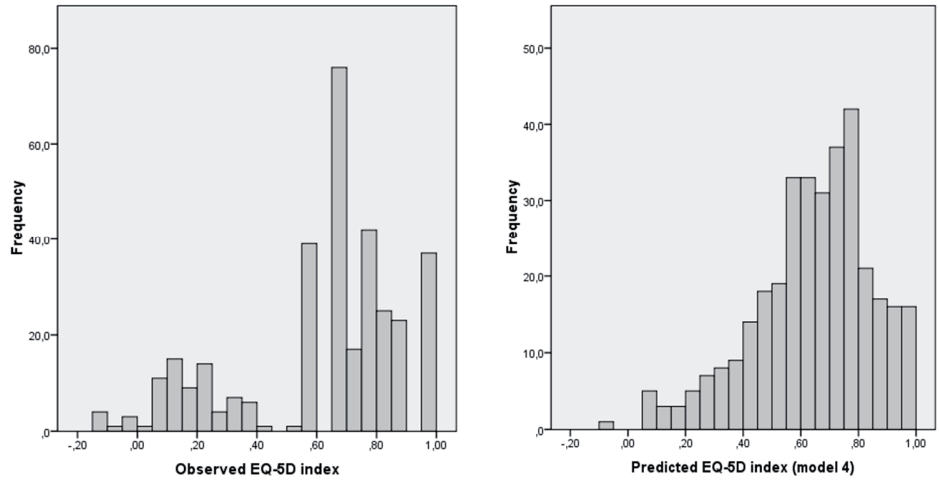


Figure 3.3 Distributions of the MSIS-29 observed and predicted EQ-5D index values

The extended model (model 3) of the HAQ can adequately discriminate between the 4 categories of disease activity. In contrast, the sum score model (model 1) does not discriminate between low and moderate disease activity. The extended model, which requires data from 3 questionnaires and as a result of missing values in one of those questionnaires, fails to predict 20 of the 123 EQ-5D index scores. The consequence is a large difference between observed (0.76) and predicted (0.83) scores for patients in the remission category due to two missing cases with lower than average utility values (both 0.67).

The MSIS-29 is a slightly different story, as the predicted EQ-5D values have more discriminative ability than the observed EQ-5D between EDSS categories 3, 4 and 5, which was noticed in both test sets and for each of the two EQ-5D country tariffs. Like the predicted EQ-5D values, the sum score of the physical impact of MS also indicates decreasing health per EDSS category. In the distribution graph (Figure 3.3), it is seen that the predicted values do not follow the distribution of the observed values (a similar pattern was observed in the cancer and arthritis data). The EQ-5D seems to have a bimodal distribution with the observations between 0.4 and 0.6 either on the low end of 0.4 or the high end of 0.5. This is most likely the result of few respondents reporting exactly those health problems on the EQ-5D which are transformed into scores between about 0.45 and 0.55 on the EQ-5D tariff.

DISCUSSION

This study aimed to develop mapping functions for HAQ, MSIS-29 and QLQ-C30. Quality of the functions was assessed with statistical indicators and performance in relevant clinical subgroups. It was also explored whether the amount of overlap between instruments could explain differences in predictive performance between mapping functions.

The best functioning mapping models are QLQ-C30 model 4, HAQ model 3 and MSIS-29 model 4. Based on the ability of the models to predict mean utility scores for the entire sample and for clinically relevant subgroups these mapping functions seem suitable for predicting utility values to rescue data for economic evaluation when a preference-based measure is absent. When correlations between the starting measure and the EQ-5D were relatively low, the mapping function performed worse. The QLQ-C30 had the highest correlation with EQ-5D dimensions and produced a function with the smallest prediction errors. The HAQ had relatively low correlations with the mobility dimension of the EQ-5D thus seemingly measuring other aspects of mobility. The content of the questionnaires reveals the differences between the measures. The HAQ sum score for mobility is made up by item 8 (walk outdoors on flat ground) and 9 (climb up 5 steps). These questions thus measure the ability to perform a specific mobility related task and differs from the interpretative EQ-5D level 2 'some problems walking about'. It may be due to this discrepancy that 114 respondents report 'no problems' on HAQ item 8, but score 'some problems walking about' on EQ-5D. The HAQ model required additional predictor variables from other questionnaires to successfully discriminate between relevant clinical subgroups.

The preferred mapping functions for the QLQ-C30, HAQ and MSIS-29 successfully predicted mean EQ-5D utility values of the test data sets. In this study, the measures of error, RMSE, MAE (and ε in the box plot), represent the average differences between all the individual predicted and observed EQ-5D index scores. The interpretation of the figures in Table 3.5 is that the mapping functions of MSIS-29 and HAQ-only models are less certain to perform well in samples that deviate 'too much' from the samples on which the models were generated or tested. All models have larger prediction errors for patients with low EQ-5D utility values, as is best represented in Figure 3.2. The functions differed in quality, despite successful prediction of the mean and relatively small prediction errors. HAQ model 1 predicts the mean correctly and has a relatively suitable RMSE, but has a small range and cannot distinguish between relevant clinical categories. Drawing on the example of the HAQ mapping function in this study, it seems that only presenting statistical measures is not always sufficient to make an educated judgment about the quality of a mapping function. Additional (clinical) indicators are more easily

interpretable and proved to be of added value for assessing the quality of the mapping functions in this study though known-groups analysis.

Improvement of the models was sought by using additional predictor variables, which were only available in the HAQ dataset. Improvement of the model was needed as HAQ model 3 (with the additional predictor variables from the SF-36, the HADS and the Disease Activity Scale) outperformed the other HAQ models in terms of RMSE, range and ability to discriminate between relevant clinical subgroups. However, here, a tradeoff is made between model performance and usability of the model, as it is not likely that many other trials included these additional predictor variables. Even in our own study the amount of missing predicted values by HAQ model 3 was much larger than for the other prediction models.

Several limitations of the study need to be discussed. First and most important, the performance of both the HAQ and MSIS-29 models is likely to be overestimated because test samples were very similar to the development samples, as they where randomly drawn from the same original dataset instead of originating from a different study, as was the case with the QLQ-C30 test set. Second, the HAQ was administered in an early arthritis cohort with relatively few very ill patients. Consequently a dummy model could not be estimated. The presence of very ill patients with low utility values is the biggest contributor to prediction errors, as these values are generally overestimated. Absence of these values in the HAQ dataset is thus likely to flatter the statistical measures of error. The extended model is recommended for use if predictor variables are available. Third, the QLQ-C30 test sample was also a lymphoma type cancer. Performance in subjects with other cancer types is not tested, which consequently causes uncertainty about model performance in different kinds of cancer. Fourth, the quality of the starting measure, the condition specific questionnaires, will have effect on the quality of the mapping function. A notable issue with the MSIS-29 mapping model was that there where many illogically ordered variables. The coefficients of scoring category 2 'a little' were higher or equal to category 3 'moderate'. A recent Rasch analysis of the MSIS-29 suggests that indeed the MSIS-29 would be better of with less answer categories (the authors of the Rasch study suggest 3), as the current 5 could not adequately distinguish the levels 'a little', 'moderately' and 'quite a bit' (101). That issue is a likely cause of the illogically ordered variables in the regression analysis.

QLQ-C30 and HAQ have been previously mapped on the UK EQ-5D tariff (90,92). The QLQ-C30 mapping, developed on a sample (N=199) of patients with inoperable esophageal cancer, did not report RMSE but reported the adjusted R^2. The adjusted R^2 was 0.61 which is somewhat lower than was found in this study, but the model could successfully predict

mean scores. We applied their mapping model, which performed well in our sample (%RMSE range over time points = 7.5-11.3). This result may provide some first support for the generalizability of QLQ-C30 mapping models, as %RMSE was relatively favorable. However, a more complete analysis involving different mapping functions tested on different types of cancer is required to investigate the important and yet unsettled issue of generalizability. The applied model, published by McKenzie and Van der Pol in 2009(90), was based on sum scores of the QLQ-C30. It is possible that our model, based on items but not all items of all QLQ-C30 dimensions, may be less sensitive in other cancer types. The previous mapping of HAQ on the EQ-5D(92) reported %RMSE of a model with dummy variables ranging from 11% to 15%, and a limited range of predicted values (0.2 -0.8). The values presented here do not deviate strongly, even though a dummy-model could not be estimated. In our study the limited range of predicted variables could be overcome by adding additional predictors that cover the other domains of general health, but this does create a problem with generalizability of the model to other datasets that do not hold all variables for instance due to the use of different instruments.

While testing the mapping functions for their discriminant validity it was noticed that MSIS-29 mapping model 4 was more sensitive than EQ-5D between categories 4 to 6 of the EDSS and had higher correlation with the EDSS. It could be argued that a change between levels 4 (able to walk 500 meters without aid) to 6 (assistance, like a cane, required to walk 100 meters without resting) on EDSS does not have impact on quality of life and is therefore not picked up by EQ-5D. However, on these same levels, there is a noticeable change on the physical sum score of MSIS-29. It was not anticipated that the predicted EQ-5D would be more sensitive to change in this area of EDSS. We hypothesize that the explanation for the difference is that the predicted and observed EQ-5D index scores have different distributions. The different distributions are most likely the result of only few respondents reporting health problems on the EQ-5D which are transformed by the country tariff into scores between about 0.4 and about 0.6 on the EQ-5D tariff (85). The predicted values of the mapping study have a different distribution than the observed values (Figure 3.3) which can be interpreted as an unsuccessful reproduction of true EQ-5D values. However, in this case, it results in an anomalous finding where a mapped EQ-5D index is more sensitive to changes in clinical categories.

The EQ-5D utility index reflects quality of life as measured on five different dimensions of health. Condition specific measures can be used to estimate mean EQ-5D utility values, despite covering different dimensions of health. However, research in this paper suggests that lower degrees of overlap leads to poorer predictive quality. Face value assessment may not necessarily represent true overlap between the dimensions of the condition specific and preference-based instrument. Mapping functions derived from

condition specific questionnaires with few dimensions may adequately predict a mean utility value, but should be used with caution in populations that deviate markedly from the population on which the function was estimated. As generalizability is a major issue for mapping functions, it ought to be tested how these models perform in different cancer, arthritis and multiple sclerosis populations. Because errors of the predicted values are larger for patients in poor health, these mapping functions may not perform well in such populations. Results from this study suggest that the mapped EQ-5D index values of the preferred mapping models can discriminate between relevant clinical subgroups. An important next step is to investigate how using mapped EQ-5D values instead of observed EQ-5D values influences cost-utility analyses.

ACKNOWLEDGEMENTS

We would like to thank Prof. J. Roberts of Sheffield University for sharing the MSIS-29 data and her useful comments on the final draft. We also thank Celina Alves, Pascal de Jong, Goedele Geuskens and Marie-Louise Lenssinck of the Erasmus University Medical Centre for preparing the HAQ data set and we thank two anonymous reviewers.

SUPPLEMENTARY MATERIAL

Table S 3.1 All developed model

	Predictors	*B*	*SE*	*p*		Predictors	*B*	*SE*	*p*		Predictors	*B*	*SE*	*p*
QLQ-C30	(Constant)	0.130	0.044	0.003	**QLQ-C30**	(Constant)	0.978	0.008	0.000	**MSIS-29**	(Constant)	0.960	0.023	0.000
Model 1	SUM_PF	0.002	0.000	0.000	**Model 4**	QLQ1	-0.030	0.010	0.002	**Model 3**	MSI3_2	-0.022	0.026	0.406
R^2 (A) = 0.71	SUM_RF	0.002	0.000	0.000	R^2 (A) = 0.74	QLQ2	-0.025	0.009	0.007	R^2 (A) = 0.57	MSI3_3	-0.042	0.027	0.124
R^2 (B) = 0.63	SUM_EF	0.003	0.000	0.000	R^2 (B) = 0.73	QLQ3	-0.045	0.010	0.000	R^2 (B) = 0.50	MSI3_4	-0.039	0.030	0.190
RMSE = 0.17	SUM_SF	0.001	0.000	0.000	RMSE = 0.13	QLQ4	-0.069	0.011	0.000	RMSE = 0.19	MSI3_5	-0.113	0.033	0.001
	SUM_GH	0.001	0.000	0.015		QLQ5	-0.159	0.016	0.000		MSI5_2	-0.035	0.022	0.115
	SUM_fatigue	0.001	0.000	0.008		QLQ6_1	-0.037	0.010	0.000		MSI5_3	-0.045	0.026	0.092
	SUM_pain	-0.001	0.000	0.000		QLQ6_2	-0.077	0.015	0.000		MSI5_4	-0.047	0.029	0.110
	SUM_const.	-0.001	0.000	0.000		QLQ6_3	-0.187	0.019	0.000		MSI5_5	-0.120	0.037	0.001
QLQ-C30	(Constant)	1.431	0.058	0.000		QLQ7_2_3	-0.020	0.011	0.084		MSI6_2	-0.049	0.027	0.072
Model 2	QLQ8	-0.033	0.012	0.006		QLQ9_1_2	-0.076	0.007	0.000		MSI6_3	-0.058	0.031	0.061
R^2 (A) = 0.74	QLQ10	-0.052	0.015	0.000		QLQ9_3	-0.267	0.019	0.000		MSI6_4	-0.053	0.033	0.109
R^2 (B) = 0.70	QLQ11	-0.040	0.015	0.009		QLQ23_1	-0.020	0.008	0.015		MSI6_5	-0.118	0.038	0.002
RMSE = 0.15	QLQ12	-0.140	0.021	0.000		QLQ23_2	-0.028	0.016	0.070		MSI10_2	-0.053	0.018	0.004
	QLQ13	-0.051	0.009	0.000		QLQ23_3	-0.267	0.048	0.000		MSI10_3	-0.034	0.022	0.126
	QLQ14	-0.015	0.009	0.093		QLQ24_1	-0.071	0.009	0.000		MSI10_4	-0.080	0.023	0.001
	QLQ16	-0.037	0.008	0.000		QLQ24_2_3	-0.144	0.015	0.000		MSI10_5	-0.075	0.030	0.014
	QLQ17	0.021	0.007	0.005		QLQ27_2	-0.041	0.010	0.000		MSI15_2	-0.029	0.020	0.159
	QLQ23	-0.031	0.010	0.001		QLQ27_3	-0.063	0.016	0.000		MSI15_3	-0.057	0.024	0.021
	QLQ26	-0.016	0.008	0.061	**HAQ**	(Constant)	0.839	0.022	0.000		MSI15_4	-0.059	0.026	0.026
	QLQ27	0.013	0.007	0.087	**Model 1**	SUM_HAQ1	-0.067	0.029	0.021		MSI15_5	-0.098	0.034	0.004
	QLQ29	-0.026	0.008	0.003	R^2 (A) = 0.39	SUM_HAQ2	0.005	0.028	0.874		MSI21_2	0.008	0.021	0.715
	QLQ30	-0.022	0.010	0.021	R^2 (B) = 0.39	SUM_HAQ3	0.002	0.025	0.929		MSI21_3	-0.022	0.023	0.334

Table S 3.1 All developed model (continued)

	Predictors	*B*	*SE*	*p*
	QLQ31	-0.051	0.011	0.000
	QLQ32	-0.016	0.008	0.039
	QLQ34	-0.036	0.007	0.000
	QLQ36	0.014	0.006	0.021
QLQ-C30	(Constant)	0.985	0.010	0.000
Model 3	QLQ1	-0.037	0.013	0.004
R^2 (A) = 0.76	QLQ2	-0.025	0.012	0.031
R^2 (B) = 0.76	QLQ3	-0.059	0.015	0.000
RMSE = 0.13	QLQ4	-0.033	0.015	0.026
	QLQ5	-0.134	0.021	0.000
	QLQ6_1	-0.033	0.015	0.035
	QLQ6_2	-0.067	0.021	0.002
	QLQ6_3	-0.180	0.027	0.000
	QLQ7_1	-0.013	0.013	0.348
	QLQ7_2	-0.037	0.019	0.054
	QLQ7_3	-0.012	0.028	0.662
	QLQ9_1	-0.065	0.010	0.000
	QLQ9_2	-0.053	0.015	0.001
	QLQ9_3	-0.189	0.031	0.000
	QLQ16_1	-0.038	0.020	0.064
	QLQ16_2	-0.045	0.033	0.181
	QLQ16_3	-0.126	0.038	0.001
	QLQ23_1	-0.028	0.011	0.011
	QLQ23_2	-0.049	0.022	0.024
	QLQ23_3	-0.456	0.108	0.000
RMSE = 0.20	SUM_HAQ4	-0.047	0.025	0.062
	SUM_HAQ5	-0.012	0.029	0.667
	SUM_HAQ6	-0.033	0.029	0.258
	SUM_HAQ7	-0.013	0.022	0.561
	SUM_HAQ8	-0.066	0.026	0.011
HAQ	(Constant)	0.836	0.020	0.000
Model 2	SUM_HAQ1	-0.070	0.026	0.008
R^2 (A) = 0.41	SUM_HAQ4	-0.050	0.022	0.028
R^2 (B) = 0.37	SUM_HAQ6	-0.040	0.025	0.104
RMSE = 0.20	SUM_HAQ8	-0.070	0.024	0.004
HAQ	(Constant)	0.527	0.113	0.000
Model 3	HAQ8	-0.038	0.024	0.106
R^2 (A) = 0.54	HADS_D	-0.017	0.005	0.001
R^2 (B) = 0.52	SF36_PF	0.001	0.001	0.107
RMSE = 0.15	SF36_RP	-0.001	0.000	0.032
	SF36_BP	0.005	0.001	0.000
	SF36_RE	0.001	0.000	0.024
	DAS28	-0.019	0.012	0.113
MSIS-29	(Constant)	1.201	0.022	0.000
	MSI21_4	-0.031	0.027	0.245
	MSI21_5	-0.129	0.033	0.000
	MSI22_2	-0.004	0.020	0.843
	MSI22_3	-0.009	0.022	0.697
	MSI22_4	-0.022	0.022	0.327
	MSI22_5	-0.112	0.029	0.000
	MSI28_2	-0.007	0.021	0.758
	MSI28_3	-0.015	0.026	0.560
	MSI28_4	-0.054	0.027	0.047
	MSI28_5	-0.056	0.032	0.080
	MSI29_2	-0.045	0.021	0.029
	MSI29_3	-0.069	0.025	0.005
	MSI29_4	-0.055	0.027	0.041
	MSI29_5	-0.194	0.031	0.000
MSIS-29	(Constant)	0.956	0.021	0.000
Model 4	MSI3_5	-0.075	0.022	0.001
R^2 (A) = 0.58	MSI5_2	-0.046	0.021	0.030
R^2 (B) = 0.49	MSI5_3	-0.055	0.025	0.029
RMSE = 0.19	MSI5_4	-0.056	0.028	0.046
	MSI5_5	-0.129	0.035	0.000
	MSI6_2	-0.062	0.025	0.013
	MSI6_3_4	-0.071	0.028	0.013
	MSI6_5	-0.134	0.036	0.000
	MSI10_2_3	-0.049	0.016	0.003
	MSI10_4	-0.084	0.023	0.000

Table S 3.1 All developed model (continued)

Predictors	*B*	*SE*	*p*		Predictors	*B*	*SE*	*p*	Predictors	*B*	*SE*	*p*
QLQ24_1	-0.053	0.011	0.000	**Model 1**	PSY_SUM	-0.011	0.001	0.000	MSI10_5	-0.082	0.030	0.006
QLQ24_2	-0.140	0.022	0.000	R^2 (A) =0.54	PHY_SUM	-0.005	0.000	0.000	MSI15_2	-0.036	0.019	0.064
QLQ24_3	-0.232	0.153	0.129	R^2 (B) = 0.51					MSI15_3	-0.068	0.023	0.004
QLQ27_1	-0.027	0.011	0.011	RMSE =0.20					MSI15_4	-0.068	0.025	0.006
QLQ27_2	-0.091	0.015	0.000	**MSIS-29**	(Constant)	1.173	0.021	0.000	MSI15_5	-0.105	0.033	0.001
QLQ27_3	-0.110	0.024	0.000	**Model 2**	MSI3	-0.022	0.008	0.005	MSI21_3	-0.031	0.018	0.084
				R^2 (A) =0.57	MSI5	-0.020	0.008	0.016	MSI21_4	-0.044	0.021	0.037
				R^2 (B) = 0 .52	MSI6	-0.020	0.008	0.019	MSI21_5	-0.141	0.029	0.000
				RMSE =0.19	MSI10	-0.024	0.006	0.000	MSI22_5	-0.098	0.024	0.000
					MSI15	-0.019	0.008	0.010	MSI28_4	-0.043	0.020	0.031
					MSI21	-0.024	0.008	0.001	MSI28_5	-0.042	0.026	0.100
					MSI22	-0.015	0.006	0.012	MSI29_2	-0.051	0.018	0.006
					MSI28	-0.018	0.007	0.018	MSI29_3_4	-0.070	0.020	0.000
					MSI29	-0.040	0.007	0.000	MSI29_5	-0.204	0.029	0.000

R^2 (A) = Adjusted R^2 in development sample

R^2 (B) = Adjusted R^2 in test sample

RMSE = Root mean square error /HADS_D= Depression sum score of HADS / SF36_PF=Physical Functioning / SF36_RP=Role-functioning / SF36BP=Bodily Pain/ SF36RE = Role-emotional / DAS28 = Disease Activity Score

Chapter 4

Condition-specific preference-based measures: benefit or burden?

With Matthijs M. Versteegh, Carin A. Uyl-de Groot and Elly A. Stolk

Published in Value in Health 2012(15): 504-513

ABSTRACT

Some argue that generic preference-based measures (PBMs) are not sensitive to certain disease specific improvements. To overcome this problem, new condition specific PBMs (CS-PBMs) are being developed, but it is not yet clear how such measures compare to existing generic PBMs.

We generated CS-PBMs from three condition specific questionnaires (Health Assessment Questionnaire for arthritis, Quality of Life Questionnaire for Cancer 30 for cancer and Muliple Sclerosis Impact Scale 29 for multiple sclerosis). First the questionnaires were reduced in content, and then, a time trade-off (TTO) study was conducted in the general public (N=402) to obtain weights associated with the dimensions and levels of the new questionnaire. Finally we compared utilities obtained using the CS-PBMs with utilities obtained by using the EuroQol five dimensional (EQ-5D) in four data sets.

Utility values generated by the CS-PBMs were higher than those of the EQ-5D questionnaire. The Health Assessment questionnaire- based measure for arthritis proved to be insensitive to comorbidities. Measures based on the Multiple Sclerosis Impact Scale 29 and the Quality of Life Questionnaire for Cancer 30 discriminated comorbidities and side-effect equally well as the EQ-5D questionnaire and were more sensitive than the EQ-5D questionnaire for mild impairments.

The introduction of PBMs which are specific to a certain disease may have the merit of sensitivity to disease-specific effects of interventions. That gain, however, is traded off to the loss of comparability of utility values and, in some cases, insensitivity to side-effects and comorbidity. The use of a CS-PBM for cost-utility analysis is warranted only under strict conditions.

INTRODUCTION

A preferred method for generating the quality adjustment required for computation of QALYs is through generic preference based measures (PBMs) such as the EuroQol five dimensional (EQ-5D) questionnaire (88) or the health utilities index (HUI) (102). Some argue that such generic PBMs are not sensitive to certain disease-specific improvements. Consequently, the existing PBMs may not always be the best tool to assess the effect of an intervention. To overcome this problem, new condition-specific PBMs (CS-PBMs) have been developed, for example, for asthma (103) and urinary incontinence (104). Not much is known, however, about how these new instruments compare with generic instruments such as the EQ-5D questionnaire. It is feared that using CS-PBMs may lead to the exaggeration of health problems due to a focusing effect, render comparison of utility values impossible, because utilities are derived from different PBMs, and may be insensitive to comorbidities (105,106). Evidence, however, is scarce. In this study, three CS-PBMs are developed for the purpose of exploring these and other issues, one for arthritis (based on the Health Assessment Questionnaire [HAQ]), one for multiple sclerosis (MS) (based on the Multiple Sclerosis Impact Scale 29 [MSIS-29]) and one for cancer (based on the EORTC Quality of Life Questionnaire C30 [QLQ-C30])

A PBM is a questionnaire with a scoring function to weight the responses according to preferences for certain health conditions over others. These preference weights are elicited in studies where respondents are asked to express their preference for a health state, for instance, using time trade-off or standard gamble. Existing generic PBMs such as the EQ-5D questionnaire and the health utilities index were developed to have a standardized tool to measure health related quality of life for the quality adjustment part of the quality adjusted life year. These generic preference-based instruments aim to measure quality of life on a sufficient degree of generality to allow comparisons across conditions. For these instruments the key tradeoff is between generality of the included health dimensions to allow cross-disease comparisons and sensitivity to pick up (relevant) treatment effects (105). The EQ-5D questionnaire, for example, consists of five items with three levels measuring mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The choice to include only these basic dimensions of health ensures the level of generality that is required for comparison across diseases at the potential cost of losing sensitivity for disease specific complaints. For example, the view is widely held that the EQ-5D is not an appropriate measure to assess quality of life of patients with sensory problems (bad eye sight or hearing problems), because sensory problems are beyond the scope of health defined by dimensions of the EQ-5D questionnaire (99). Another perceived problem of EQ-5D is that very mild conditions cannot be adequately assessed using only three levels of impairment due to low ceiling sensitivity (107,108).

The increased use of economic evaluations by health authorities seems to have created a sense of urgency within the health assessment community to deal with the shortcomings of generic PBMs. In recent years new CS-PBMs have emerged for which the development was motivated by either the absence of generic PBMs in a specific context or the judgment that generic PBMs would not be appropriate for a condition. Contrary to generic instruments, a CS-PBM contains dimensions specifically targeted at the affected population. In terms of the trade-off mentioned above, these instruments are expected to demonstrate superior sensitivity to specific diseases, although this may come at the cost of comparability of utility values across conditions. Because of the difference in the scope of different instruments, utility values derived from a CS-PBM may not be comparable with those derived from a generic instrument, even though they seem to lie on the same 0 to 1 scale. Although the development of CS-PBMs is valuable for research purpose, for example to better investigate the shortcomings of generic PBMs, there is concern about the application of CS-PBMs in economic evaluations. Unfortunately, empirically founded guidance on how and when to apply CS-PBMs is absent.

There has been little reflection so far on the comparability of the obtained quality of life weights to those obtained from generic PBMs. Specific issues in comparability are described in a recent expert editorial (105). First, CS-PBMs may cause an exaggeration of health problems (reflected by low utility values) due to focusing effects. When the health states in a preference elicitation study consist of a set of disease-related items, rather than general items of health-related quality of life, the context of the valuation is narrower, possibly leading to lower utility values. The logic behind this hypothesis is that narrow focused items may seem less important when presented in a wider context of general health (e.g., having a cold may seem less severe when presented alongside problems with mobility), but may seem quite problematic when presented separately. This may result in a downward bias on preference values when compared to generic PBMs. Second, a CS-PBM might have difficulty capturing comorbidities as the focus is on disease-related items. This may result in an upward bias on utility values. Furthermore, developing a CS-PBM is not a clear cut exercise. Researchers face many decisions, such as the reduction of items in a questionnaire, the selection of health states (how many and which?) that have to be valued to develop a PBM (99), on the valuation method (e.g., TTO or standard gamble?), and on the modeling approach. How these decisions are dealt with may differ per study which decreases comparability.

The primary aim of this paper is to provide empirical evidence about the comparability of CS-PBMs and generic PBMs. To do so, three CS-PBMs were developed from existing questionnaires. The values generated by these CS-PBMs were then compared with EQ-5D questionnaire values for the same patient samples. By providing empirical evidence we

hope to provide a better understanding of the effects of using CS-PBMs and contribute to development of guidance for their use. This is important, as it can be expected that in the nearby future more cost utility analyses will contain utilities based on conditions specific measures.

METHODS

Questionnaires for CS-PBM development

The CS-PBMs were developed from the HAQ (91), the MSIS-29 (93) and the QLQ-C30 (89). These instruments were selected based on expert advice and commonality of use within clinical settings. The HAQ is a widely used questionnaire in rheumatology to measure functional abilities using 20 items with four levels spread across eight domains (dressing, rising, eating, walking, hygiene, reach, grip and usual activities). The scale has been shown to be unidimensional (109). The MSIS-29 measures the impact of MS on a physical and psychological dimension. Dimensionality of the subscales has been confirmed using Rasch analysis (101). The QLQ-C30 (version 2) is a cancer specific questionnaire consisting of 30 items. These items cover five functional scales, nine symptom scales and a global health status scale. These questionnaires were chosen because they differ in scope and because EQ-5D data was available for the purpose of comparing results. For MSIS-29, it has been shown that the physical scale is better capable of discriminating among sub categories of the clinically assessed Expanded Disability Status Scale (EDSS) than is the EQ-5D questionnaire (110). There was no evidence known to us on a lack of responsiveness of EQ-5D or the superiority of the condition-specific measures HAQ and QLQ-C30 in arthritis or cancer, respectively.

Reducing the content of the questionnaires

Developing a PBM from an existing questionnaire does not lead to an entirely new instrument but attaches weights to some of the items of the existing questionnaire. Such an approach generally requires a method to reduce the questionnaire content as only a limited number of items can be valued in a preference elicitation study (99). Typically only a fraction of the total amount of all theoretically possible health states is valued. The values for the remainder of the health states are estimated through modeling techniques.

The optimal number of items in a health state was considered to be in the order of five to nine items, because more items may cause difficulties for respondents in the valuation study (99). The HAQ, MSIS-29 and QLQ-C30, respectively, contain 20, 29 and 30 items, so reduction of content was required. Relevant and well-functioning items from the

questionnaires were selected using the following criteria proposed by others (111): *i)* the item had to fit the Rasch model, *ii)* the item had to meet basic psychometric criteria and *iii)* the selected items had to be approved by a clinical expert. Four datasets were available for these analyses: the Rotterdam Early Arthritis CoHort for the HAQ (N=738), the Multiple Sclerosis Risk Sharing Scheme Monitoring Study (N=1,295) for the MSIS-29 and the Hemato Oncology Fouccation for Adults in the Netherlands 24 (pooled N=716) and Hemato Oncology Fouccation for Adults in the Netherlands 25 (pooled N=789) trials for the QLQ-C30. The dataset characteristics are described in detail in Versteegh et al. (2011) (110). A set of a priori criteria were used to determine which items were suitable for the health state description (111,112). Because it was expected that neither of these criteria could be sufficient on its own, the three criteria were employed 'side by side' (i.e. no hierarchical order).

Criterion 1: fit to the Rasch model

Rasch analysis was used to test the psychometric validity of a scale and to identify well-functioning items. The Rasch model assumes that the probability of scoring level λ on item *i* is a logistic function of the relative distance between the item location (how much disability it represents) and the respondent location (how disabled the patient is) (113).

The main performance criteria within the Rasch model were whether the item: *i*) has ordered thresholds (having more of the latent trait θ results in endorsing a higher level answer category (114)); *ii*) fits the Rasch model (fit residual <2.5 and non-significant bonferroni adjusted probability); *iii*) combined scale fits the Rasch model (described by a non-significant item-trait interaction chi-square probability (114)) and *iv*) shows no differential item functioning. After each single scale amendment the analysis was rerun for the remaining items. Rasch analysis was performed on the dimensional structure originally suggested by the questionnaires.

Criterion 2: psychometric properties

Psychometric criteria were laid alongside the Rasch results to come to a final selection of items amenable for valuation. The functioning of the items was tested by investigating the loading of items on factors identified by factor analysis; missing data; internal consistency of items with its scale score; distribution of the responses on an item including floor and ceiling effects; and regression coefficients between a general health indicator and an item. Psychometric analyses were applied to the full data sets.

Criterion 3: expert opinion

The selected items from the questionnaires were presented to experts in the respective fields. Experts from the Erasmus Medical Centre and the VU Amsterdam Medical Centre

were consulted to gain insight in important aspects of the disease and to evaluate the result of the previous selection process.

Health state selection

Even after data reduction the selected set can still generate an enormous amount of possible health states; therefore, a fractional factorial design was favored over a full factorial design. The QLQ design was a level-balanced design, meaning that all levels of each item occurred with the same frequency. Within the balanced design health states covered the entire spectrum of severity, measured by averaging the item levels of a health state. For the MSIS-29 and the HAQ, items and levels were selected with an orthogonal main effects plan (OMEP) as is applied in other studies (83,115) to ensure zero statistical correlation between the attributes. The set was complemented with a selection of the most observed health states (four or more observations) over the severity range of the questionnaire. TTO values estimated with additive main-effect models (one based on the OMEP states and one based on the OMEP and the most observed states) were compared to the observed TTO values of the most occurring states using standard predictive performance measures like mean absolute error (MAE) to see if the addition of these states led to improved prediction of the most frequently occurring states.

The final design was blocked. In such a design respondents value a number of health states which belong to the same 'block'. The mean severity of the combination of items in a block was similar and measured through summing the level scores of the items in a block.

Health state valuation with Time trade-off method

The preferences of a sample of the general public were elicited through a TTO exercise for each of the selected health states of the questionnaires. To optimize comparability to generic PBMs the CS-PBM health states were valued with the same TTO protocol, the same computer assisted personal interview tool, the same procedure to measure states 'worse than dead' and the same rescore procedure of negative values (negative TTO scores were rescaled to have a range between -1 and 0 with (-t/-x-t) as was adopted in the Dutch EQ-5D valuation study (87). Unlike the Dutch EQ-5D valuation study, this study was performed in group sessions, which has previously been shown to produce comparable TTO results (116).

The TTO exercise was self-administered through a digital tool for TTO elicitation (computer assisted personal interviews) in groups with about 12 to 25 respondents per session. Each session was supervised by three to four researchers to offer assistance if needed. Prior to the task, respondents received 30 minutes of instructions by research-

ers M.V. or A.L. including examples of the TTO computer program projected on a large screen. The task was piloted by M.V. and A.L. in a sample of 18 respondents to ensure the introduction, the computer program and the organization of the task were feasible.

The three questionnaires were presented separately in the TTO exercise and in all possible orders (e.g. first HAQ, then MSIS then QLQ). Within the TTO exercises, health states were presented random to individuals.

Respondents

Respondents were selected by a marketing agency which requited a sample resembling the Dutch general population in age, gender and education. Respondents were approached by phone and asked if they were interested in contributing to a task to value descriptions of health states. Respondents received a financial reward of € 35, - upon completion of the three TTO exercises. Respondents were removed from the analyses when the results indicated they valued the majority of logically worse states higher than logically better states in a set (i.e. HAQ state 11112 is logically better than HAQ state 14444).

Modeling of the TTO values

Once the TTO study had been performed, the preference values observed for the selected health states were used to estimate values for all potential health states through statistical modeling. Because individuals value more than one health state there are multiple observations for each individual. Random effects models were estimated to assess how the predictors (the items and their levels) influence the dependent variable (the mean observed TTO value). In these random effects models, the item levels were treated as dummy variables with dummy coding. The constant term was treated as an additional decrement for having any item level other than the base case ('no problems'), which is similar to the EQ-5D model. The values predicted by this random effects model will be referred to as the PBM results (e.g. HAQ-PBM). Models were required to have significant predictors and worse scores on the levels ought to be represented by larger utility decrements. Model performance was assessed by comparing the MAE of observed and predicted values. Models were estimated until meeting those criteria. Only the most parsimonious models are presented. To keep optimal comparability between the developed CS-PBMs, models were estimated from the items only, without interaction effects or a 'worst-value' dummy variable which is 1 for every item on the lowest level. Interaction effects were not estimated because the study design was a main effects design.

Hypotheses and analyses

To investigate the properties of preference-based measures developed from existing questionnaires several hypotheses were tested. First, it was tested whether the TTO values could be successfully modeled. For HAQ and MSIS the TTO random effects models were fitted on both the full dataset (including 'most observed' health states) and on the subset consisting of health states originating from the OMEP. This was done to test whether an OMEP alone is sufficient to estimate the utility values of the most occurring health states. Second, it was investigated whether CS-PBMs yielded lower mean utility values than a generic measure, which was hypothesized to reflect that a downward bias on utility values resulting from a focusing effect might outweigh the upward bias on utility values resulting from a narrower scope of the CS-PBM. Third, it was tested with Wilcoxon rank-sum tests, to account for the non-normal distribution of utility values, whether the developed CS-PBMs had a more narrow focus and were therefore less sensitive to comorbidities (in arthritis and MS datasets) or side-effects (Non-Hodgkin's lymphoma dataset) than EQ-5D. Side-effects had World Health Organization performance status 2 or higher, representing the inability to carry out work activities due to the condition. Fourth, we assessed discriminative properties of the new measures by using clinical indicators. For arthritis the Disease Activity Score-28 was used, which is based on a count of tender joints and the erythrocyte sedimentation rate. It distinguished between high, moderate and low disease activity, and remission. For MS the EDSS was used which, when rounded to integers, distinguishes 11 categories of increasing disability. For cancer we used the WHO performance status score (or Eastern Cooperative Oncology Group) which distinguishes 6 categories, from 0 to 6 (death). Lastly, responsiveness was measured in the cancer population using effect-size (Cohen's *d*) and mean change in the cancer population, for which follow-up measurements were available in the data set.

All results were compared to utilities of the Dutch EQ-5D tariff (22).

Software

For Rasch analysis the RUMM2020 software (Rumm Laboratory Pty Ltd) was used. Psychometric analysis was performed in SPSS 17.0 (SPSS Inc.) and all hypothesis testing and modeling efforts in STATA 11.0 (StataCorp. 2009).

RESULTS

Item and level selection

The selected items per questionnaire are presented in Table 4.1, and all met the criteria of the Rasch analysis, psychometric analysis and expert opinion. The full results of the selection of items and levels and the results of the Rasch analysis are presented in the supplementary material.

Table 4.1 Items selected for the TTO valuation exercise

HAQ-DI	MSIS-29	QLQ-C30
• HAQ1 Stand up from a straight chair • HAQ2 Walk outdoors on flat ground • HAQ3 Get on / off toilet • HAQ4 Reach and get down a 5-pound object (such as a bag of sugar) from just above your head • HAQ5 Open car doors	• MSIS1 Problems with your balance • MSIS2 Being clumsy • MSIS3 Limitations in your social and leisure activities at home • MSIS4 Difficulties using your hands in everyday tasks • MSIS5 Having to cut down the amount of time you spent on work or other daily activities • MSIS6 Feeling mentally fatigued • MSIS7 Feeling irritable, impatient or short tempered • MSIS8 Problems concentrating	• QLQ1 Trouble taking a long walk • QLQ2 Limited in doing either your work or other daily activities • QLQ3 Have you had pain • QLQ4 Have you felt nauseated • QLQ5 Were you tired • QLQ6 Difficulty in concentrating on things • QLQ7 Did you worry • QLQ8 Has your physical condition or medical treatment interfered with your social activities

Resulting study design

Given the many items and levels in the study we chose a fractional factorial blocked design. Health states were presented in blocks, so one individual values one block containing several health states. The design is summarized in Table 4.2.

Table 4.2 TTO study design following item selection

	HAQ	MSIS	QLQ
Number of items	5	8	8
Total amount of health states to be valued	56	100	105
States identified by OMEP (used in study after fold-over)	15 (30)	32 (64)	n/a
Number of most occuring states included	26	36	n/a
Number of states valued by one individual (total = 33)	8	10	15
Number of blocks[1]	7	10	7
Expected number of observations per health state (N=400 / number of blocks)	57	40	57

[1] One block consist of a number of states and all of the states in one block are valued by one individual

Data quality

Four hundred two respondents participated in the computer assisted TTO study and resembled the Dutch population (Table 4.3). Respondents were excluded from the analyses because they had valued the majority of logically better states lower than logically worse states in one block (8 exclusions for HAQ, 17 for MSIS and 7 for QLQ). Average time to value one health state in the TTO exercise was about 1 minute. Total time per block was highest for QLQ (15 health states, about 12 minutes), followed by MSIS (10 health states, 10 minutes) and HAQ (8 health states, 8 minutes). Although two separate researchers took turns in holding the introductory talks this did not bias the TTO responses (Wilcoxon rank test p>0.05). On average, women had higher utility values (t-test, $p<0.00$) for all three questionnaires. The mean utility of the health states and the percentage of responses indicating a state to be worse than dead are presented in Figure 4.1A and B.

Table 4.3 Respondent characteristics

	TTO study sample	Dutch population norms[1]
N	402	-
Gender M/F	46% / 54%	49.5% / 50.5%
Age		
mean (SD)	45 (15.5)	40.1
min-max	15-76	-
Agegroup		
<20	4.8	23.7
20-40	37.6	25.3
40-65	46.9	35.7
65-80	10.4	11.4
>80	0.3	3.9
Education		
High	34	27
Medium	35	31
Low	25	33
Missing / Else	6%	9%
Mean (SD) time to complete TTO		
HAQ 8 states	8 min (4.6min)	-
MSIS 10 states	10 min (5.8min)	-
QLQ 15 states	12.7 min (5.8min)	-

[1] Statistics Netherlands, 2009 figures.

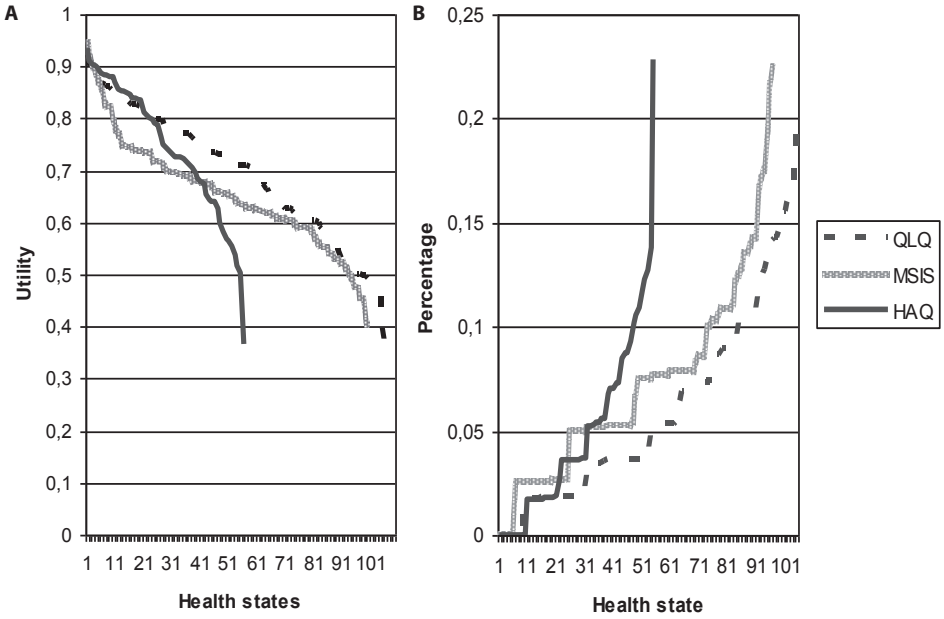


Figure 4.1 A. Mean utility values of health states, B. Percentage of respondents who classified a state worse than dead

MSIS = Multiple Sclerosis Impact Scale, HAQ = Health Assessment Questionnaire, QLQ = Quality of life questionnaire,

Modeling

TTO values were modeled for each of the three questionnaires with random effects mean prediction models. For the HAQ, using only the OMEP based health states had too much variation in TTO scores between respondents to estimate a model with significant predictors and logical negative signs for each of the dummy variables (increasing negative decrements per item level of severity). Estimating the model on all the available data (thus including the 'most observed' states) yielded a well-functioning

Table 4.4 Final random effects model characteristics

	HAQ-PBM†	MSIS-PBM*	MSIS-PBM†	QLQ-PBM
Random effects mean models				
R†	0.94	0.68	0.78	0.88
MAE	0.028	0.034	0.04	0.033
MAE most observed states	0.022	0.057	0.043	-
Illogical sign or order of variables	0	0	0	0
Insignificant predictors	0	0	0	0
Possible range	0.32 - 1	0.40 - 1	0.42 - 1	0.34 - 1

* Model based on states from the orthogonal design

† Model based on states from the orthogonal design and the most observed states

mean prediction model. The prefinal MSIS-29 model had insignificant predictors for three variables MSIS3; the pre-final QLQ-C30 model had insignificant predictors for two variables. In all instances merging the levels with the adjacent categories resolved the problem. Model characteristics are summarized in Table 4.4 and full models are presented in Table 4.5.

When the MSIS-29 prediction model was based on all the states (thus including the most observed states), the prediction error for the most observed states was reduced (MAE=0.043 compared to MAE=0.057). When the MSIS-29 TTO values were modeled without the 'most observed' states the utility values were generally higher which caused that the utility values of some of the 'most observed' states were overestimated (Figure 4.2).

Table 4.5 Coefficients of random effects models with TTO value as dependent variable

HAQ-PBM			MSIS-PMB†			QLQ-PMB		
	Coefficient	Std. Err.		Coefficient	Std. Err.		Coefficient	Std. Err.
haq1_2	-0.005	0.001	ms1_2	-0.016	0.003	qlq1_2	-0.027	0.001
haq1_3	-0.031	0.002	ms1_3	-0.043	0.003	qlq2_2	-0.020	0.002
haq1_4	-0.121	0.002	ms1_4	-0.089	0.003	qlq2_3	-0.047	0.002
haq2_2	-0.029	0.001	ms2_2	-0.018	0.003	qlq2_4	-0.068	0.002
haq2_3	-0.091	0.002	ms2_3	-0.047	0.003	qlq3_3	-0.079	0.002
haq2_4	-0.144	0.002	ms2_4	-0.047	0.003	qlq3_4	-0.213	0.001
haq3_2	-0.042	0.001	ms3_3	-0.055	0.002	qlq4_2	-0.018	0.002
haq3_3	-0.055	0.002	ms3_4	-0.071	0.002	qlq4_3	-0.055	0.002
haq3_4	-0.213	0.002	ms4_2	-0.061	0.002	qlq4_4	-0.089	0.002
haq4_2	-0.022	0.001	ms4_3	-0.101	0.003	qlq5_2	-0.021	0.002
haq4_3	-0.041	0.002	ms4_4	-0.108	0.003	qlq5_3	-0.031	0.002
haq4_4	-0.074	0.002	ms5_2	-0.032	0.003	qlq5_4	-0.037	0.002
haq5_2	-0.016	0.001	ms5_3_4*	-0.057	0.002	qlq6_2	-0.004	0.002
haq5_3	-0.038	0.002	ms6_2	-0.020	0.003	qlq6_3	-0.039	0.002
haq5_4	-0.044	0.002	ms6_3	-0.035	0.003	qlq6_4	-0.052	0.002
Constant	0.918	0.002	ms6_4	-0.059	0.003	qlq7_3	-0.009	0.002
			ms7_3	-0.024	0.002	qlq7_4	-0.047	0.002
			ms7_4	-0.038	0.002	qlq8_2	-0.008	0.002
			ms8_2	-0.037	0.003	qlq8_3	-0.041	0.002
			ms8_3	-0.059	0.003	qlq8_4	-0.060	0.002
			ms8_4	-0.073	0.003	Constant	0.944	0.002
			Constant	0.959	0.005			

* Both ms5_3 and ms5_4 have the same decrement

† Msis model with most observed health states included

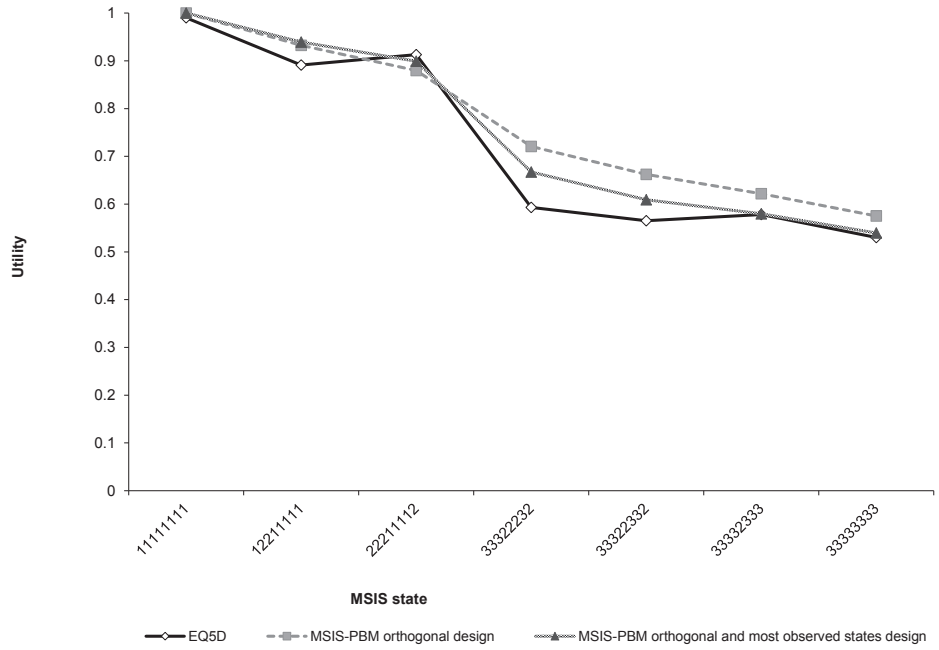


Figure 4.2 Utility values for most observed MSIS states in patient dataset

Table 4.6 Comparison of utility values derived from the new PBM measures with EQ-5D and SF-6D

	HAQ	MSIS	QLQ_MM‡	QLQ_NH‡
N	738	1,295	716	789
Mean utility (SD) [range]				
EQ5D	0.68 (0.23) [-0.134 - 1]	0.62 (0.26) [-0.220 - 1]	0.74 (0.21) [-0.058 - 1]	0.73 (0.26) [-0.330 - 1]
SF6D	0.66 (0.10) [0.370 - 1]	-	-	-
PBM*	-	0.69 (0.13) [0.400 - 1]	0.84 (0.09) [0.440 - 1]	0.82 (0.11) [0.340 - 1]
PBM†	0.91 (0.09) [0.570 - 1]	0.67 (0.14) [0.420 - 1]	-	
Intraclass correlations				
EQ5D-PBM*†	0.45	0.62	0.64	0.67

* Model based on states from the orthogonal or balanced design
† Model based on states from the orthogonal design and the most observed states
‡ MM= Multiple Myeloma, NH = Non-Hodgkin

Comparability of mean utility values

In the four datasets, the developed CS-PBMs based on the models presented in Table 4.5 produced higher mean utility score for patients than did the EQ-5D questionnaire (Table 4.6). Especially the HAQ-PBM (mean = 0.91) had a much higher mean utility value than did the EQ-5D questionnaire (mean = 0.68). Furthermore, the difference between the mean EQ-5D questionnaire score in arthritis and the mean EQ-5D questionnaire score in

Table 4.7 MSIS-PBM and QLQ-PBM have increased sensitivity at the ceiling of EQ-5D

Total sample size		EQ5D =1	EQ5D <1	Worst state for which EQ-5D = 1
738	HAQ-PBM < 1	n = 7	-	21211
	HAQ-PBM = 1	-	n = 252	
1,295	MSIS-PBM < 1	n = 99	-	33111222
	MSIS-PBM = 1	-	n = 2	
1,505	QLQ-PBM < 1	n = 185	-	24334324
	QLQ-PBM = 1	-	n = 4	

MS was 0.06 while the differences between the HAQ-PBM and the MSIS-PBM are 0.24. The QLQ-C30-PBM based utility values had the highest correlation with EQ-5D utility values. Both the MSIS-PBM and the QLQ-PBM, however, have increased sensitivity compared to EQ-5D. Where EQ-5D scores full health (a utility value of 1) the MSIS-PBM and the QLQ-PBM report decrements in utility for respectively 99 and 185 patients (Table 4.7).

Comorbidities and side-effects

The HAQ-PBM could not discriminate between patients with and without comorbidity (other vascular disorders and psychiatric disorders) when EQ-5D could (Table 4.8). For arthritis patients with diabetes, hypercholesterolemia or thyroid disease the HAQ-PBM showed higher utility values for individuals with the disorder while EQ-5D signaled the expected direction of differences. The MSIS-PBM also showed higher utilities for patients with asthma and high blood pressure (rather than without) but this was concordant with the differences indicated by the EQ-5D questionnaire. Both the MSIS-PBM and the EQ5D picked up significant differences between MS patients with and without depression ($p<0.05$). In the non-Hodgkin's lymphoma dataset, patients with side-effects and infections as result of treatment had lower ($p<0.05$) utility values in both EQ-5D and QLQ-C30 than patient without side-effects and infections, except for hair loss. All significant differences were at least half a SD except for comorbidity 'depression' in the MS dataset and 'other side-effects' in the non-Hodgkin's lymphoma dataset.

Discriminative ability and responsiveness

Utilities of all instruments decreased with an increase of severity as assessed by the clinical indicator (Table 4.9). The utilities of the HAQ-PBM, however, failed to distinguish between low and moderate disease activities. EQ-5D did so accurately. As has previously been shown, EQ-5D was unable to distinguish between categories 3, 4 and 5 on EDSS (15). This signifies the inability of EQ-5D to distinguish between fully ambulatory MS patients (EDSS 3) and patients whose disability is severe enough to impair full daily working activities (EDSS 5). The MSIS-PBM, of which the physical scale was known to be sensitive to changes between level 3 4 and 5, did pick up the deterioration in health.

Neither the QLQ-PBM nor the EQ-5D adequately reflected the deterioration between level 0 and level 1 of the WHO performance status.

The QLQ-C30 was, in terms of effect-size measured with Cohen's *d*, at times more, and at times less sensitive to changes over time (Table 4.10). However, the absolute differences indicated that even when the QLQ-PBM had a larger mean difference relative to the standard deviation, the EQ-5D still reported larger mean change scores.

Table 4.8 comorbidities and side-effects by PBM and EQ-5D

	HAQ-PBM				EQ5D			
	Comorbidity		No comorbidity		Comorbidity		No comorbidity	
	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median
Diabetes	0.92 (0.08)	0.90	0.89 (0.10)	0.89	0.66 (0.24)	0.79	0.71 (0.16)	0.78
Hyper-cholesterolemia	0.90 (0.09)	0.88	0.89 (0.10)	0.89	0.57 (0.27)	0.65	0.72 (0.15)	0.78
Thyroid disease	0.90 (0.10)	0.89	0.89 (0.10)	0.89	0.60 (0.29)	0.71	0.72 (0.16)	0.78
Other cardiac disease	0.86 (0.05)	0.88	0.89 (0.10)	0.89	0.60 (0.21)	0.68*	0.72 (0.15)	0.78*
Psychiatric disorder	0.84 (0.13)	0.89	0.89 (0.10)	0.89	0.54 (0.30)	0.67*	0.72 (0.16)	0.78*
	MSIS-PBM				EQ5D			
	Comorbidity		No comorbidity		Comorbidity		No comorbidity	
	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median
Depression	0.61 (0.12)	0.59*	0.68 (0.14)	0.68*	0.54 (0.26)	0.64*	0.63 (0.26)	0.67*
Asthma	0.67 (0.14)	0.71	0.67 (0.14)	0.68	0.63 (0.22)	0.68	0.62 (0.26)	0.67
HPB	0.68 (0.14)	0.73	0.65 (0.13)	0.68	0.64 (0.25)	0.64	0.60 (0.26)	0.67
	QLQ-PBM				EQ5D			
	Side-effects		No side-effects		Side-effects		No side-effects	
	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median
Neurotoxicity†	0.76 (0.09)	0.76*	0.81 (0.09)	0.83*	0.56 (0.27)	0.65*	0.73 (0.25)	0.81*
Hair loss†	0.80 (0.08)	0.79	0.81 (0.10)	0.83	0.70 (0.23)	0.78	0.72 (0.26)	0.79
Nausea†	0.73 (0.13)	0.74*	0.81 (0.09)	0.83*	0.56 (0.31)	0.65*	0.72 (0.25)	0.81*
Other side-effects†	0.79 (0.09)	0.80*	0.81 (0.09)	0.83*	0.66 (0.23)	0.69*	0.72 (0.26)	0.81*
	Infection		No infection		Infection		No infection	
Ear / Nose / Throat[1]	0.72 (0.11)	0.72*	0.81 (0.09)	0.83*	0.42 (0.32)	0.31*	0.73 (0.25)	0.81*

* = significant difference between comorbidities / no comorbidities at ($p<.05$) Wilcoxon sum rank test
† WHO grade ≥ 2

Table 4.9 Discriminant validity

	HAQ-PBM		EQ-5D		
DAS28	Mean	SD	Mean	SD	N
Remission	0.98	0.04	0.76	0.20	11
Low DA	0.90	0.08	0.70	0.25	15
Moderate DA	0.90	0.09	0.67	0.22	70
High DA	0.83	0.07	0.51	0.29	27
	MSIS-PBM		EQ-5D		
EDSS	Mean	SD	Mean	SD	N
0	0.80	0.14	0.81	0.22	35
1	0.78	0.14	0.78	0.23	74
2	0.73	0.14	0.72	0.23	262
3	0.68	0.14	0.63	0.25	206
4	0.66	0.13	0.63	0.23	248
5	0.63	0.10	0.64	0.19	103
6	0.60	0.11	0.54	0.25	201
7	0.58	0.11	0.46	0.27	78
8	0.57	0.07	0.40	0.31	17
9	0.47	0.07	0.09	0.10	5
	QLQ-PBM		EQ-5D		
WHO	Mean	SD	Mean	SD	N
0	0.83	0.11	0.75	0.25	356
1	0.83	0.10	0.76	0.24	304
2	0.80	0.11	0.69	0.24	96
3	0.71	0.10	0.37	0.27	27

DAS-28 = Disease-activity score-28, EDSS = Expanded Disability Status Scale, WHO = World Health Organization

Table 4.10 Responsiveness of utilities in Non-Hodgkin sample

	Cohen's d		Mean change	
Follow-up	QLQ-PBM	EQ-5D	QLQ-PBM	EQ-5D
2nd treatment cycle	0.13	0.17	0.02	0.05
4th treatment cycle	0.02	0.08	0.00	0.02
6th treatment cycle	-0.09	-0.06	-0.01	-0.01
3 months follow-up	0.33	0.22	0.03	0.06
6 months follow-up	0.25	0.10	0.02	0.02
10 months follow-up	-0.01	-0.09	0.00	-0.02
18 months follow-up	0.00	0.19	0.00	0.04

DISCUSSION

This study developed three CS-PBMs from existing questionnaires HAQ, MSIS-29 and QLQ-C30 to provide evidence concerning comparability of CS-PBM derived utility values with generic PBM-derived utility values. CS-PBMs had different mean utility values within a disease and did not report equal differences in mean utility values between diseases. The CS-PBMs in this study did not seem to exaggerate health problems, but rather reported higher mean values. Capturing comorbidities and along that line: side-effects of interventions appeared problematic for the HAQ-PBM, but not for MSIS-PBM and QLQ-PBM. The MSIS-PBM and QLQ-PBM were more sensitive than the EQ-5D questionnaire to very mild impairments. The physical scale of the MSIS-29 questionnaire is known to be more sensitive in discriminating between clinical categories in multiple sclerosis than is the EQ-5D questionnaire. The MSIS-PBM, derived from the MSIS-29, also has better discriminatory properties.

Because the mean utility values of all three CS-PBMs were higher than those of generic instruments, it seems that a potential downward bias of a focusing effect may be smaller in size than the upward bias that results from a narrower scope of the condition specific measures. This is most clearly seen in the performance of the HAQ-PBM, which is developed from the HAQ-Disability Index which measures functional ability (91). Consequently the HAQ-PBM indicates the utility decrements associated with these functional (dis)abilities. In the HAQ-PBM there is no dimension such as 'pain' or 'psychological state'. Because pain is a frequently occurring symptom in arthritis, it is not surprising that the mean utility value of the early arthritis cohort as measured by the HAQ-PBM is much higher than the mean utility value of the generic instruments; any additional utility decrement besides functional disabilities, such as pain, is not captured directly, if at all. In the case of the HAQ, this result could have been anticipated based on the fact that the HAQ-Disability index aims to offer a unidimensional assessment of functionalities and does not attempt to measure other dimensions of health since these are captured by other instruments that are part of the minimum dataset internationally agreed on. The unidimensionality of the HAQ caused some problems in the valuation task. Because all items aim to measure the same underlying latent variable (functional ability) they are highly related. OMEP generated states have favorable statistical properties, but do not consider the sensibility of the combination of item levels. Consequently, one health state in the valuation study consisted of the counterintuitive combination "able to get up from a chair" and "not able to get up from the toilet". This particular state caused confusion with some of the respondents.

The HAQ-Disability index does not intend to form a comprehensive assessment of relevant disease specific health outcomes in patients with rheumatoid arthritis, and therefore could be rejected as offering a suitable basis for development of CS-PBMs. The large deviations in mean utility values presented in this study between the HAQ-PBM and EQ-5D support this view. More generally, it can be concluded that instruments with a narrow scope, often identifiable through inspecting items or dimensions, are unsuitable as a base for CS-PBMs used for resource allocation.

The perceived insensitivity of existing generic instruments is an important motive for developing CS-PBMs. In this study sensitivity of the CS-PBM and EQ-5D was compared by investigating ceiling effects and discriminative ability of the instruments between patients with and without comorbidity or side-effects. A ceiling effect found in EQ-5D for mild impairments was not found in the MSIS-PBM and QLQ-PBM (Table 4.7). One reason for this difference may be the descriptive system of the questionnaires: the three-level system of EQ-5D might result in a lower likelihood of reporting problems than the four-level systems of the CS-PBMs. Nevertheless, using CS-PBMs did not result in an exaggeration of health problems on average when compared to generic instruments in this study. Rather, the mean utility value of MSIS-PBM and QLQ-PBM was higher than EQ-5D. This may be a reflection of the smaller range in obtainable utility values, which skews the average upwards. Bad EQ-5D health states reflect very poor health, which is perhaps not captured in MSIS-PBM and QLQ-PBM. Indeed, the negative range of utility values as produced for EQ-5D has rarely been reproduced for other instruments. EQ-5D, MSIS-PBM and QLQ-PBM performed equally well in distinguishing patients with comorbidities / side-effects from patients without it. Only the HAQ-PBM performed poorly in this aspect. Interestingly, the MSIS-PBM and the QLQ-PBM displayed equal discriminative ability as EQ-5D despite having a much smaller total scale size due to a higher 'floor' (i.e. the lowest attainable value).

Superiority of CS-PBMs compared to EQ-5D in regard to their discriminative ability is not demonstrated for HAQ-PBM and equivalence has been shown for QLQ-PBM. MSIS-PBM showed better discriminative properties than did the EQ-5D questionnaire in EDSS subcategories. With additional evidence on known-group differences this could prove the MSIS-PBM to be a contribution to cost-utility analyses. The original preference-based questionnaire MSIS-29 was the only measure for which empirical evidence indicated better discriminative properties than EQ-5D in a multiple sclerosis data sets.

While a CS-PBM may have desirable statistical properties, such as expressed in effect-size or the ability to identify significant differences between groups with or without side effects, partly due to a small SD of mean values, these properties may not be reflected

the absolute size of differences in utility values between groups. This has consequences for QALY computation. Imagine a new drug that reduces nausea from cancer treatments. Using the figures from Table 4.8, the population not having nausea would have a higher utility with an effect-size (Cohen's *d* with pooled SDs) of 0.57 for EQ-5D but a larger 0.73 for QLQ-PBM. The absolute difference, however, would be 0.16 for EQ-5D and 0.08 for QLQ-PBM. An implication of these results is that if a CS-PBM is developed in order to increase sensitivity compared to EQ-5D, statistical sensitivity is not a sufficient criterion.

Rather than due to concerns about the sensitivity of an existing generic PBM, a CS-PBM may also be developed because a PBM was not administered in, for example, a clinical trial. In this case one could also choose to use the variation in responses on a condition-specific measure to estimate what a generic utility instrument like EQ-5D would have been had it not been absent, a process called mapping (32). It is important to reflect on the question which strategy for deriving utilities from a disease-specific instrument is most appropriate. The main difference between mapping and constructing a CS-PBM is that the development of a PBM assigns population weights (via TTO) to the item levels of a questionnaire, while a mapping function assigns weights to the items that are dependent on the generic measure it aims to estimate. As such, issues with insensitivity of the generic instrument are not resolved when mapping a condition specific measure onto a generic PBM. In our view, a well conducted and validated mapping function may be preferred to the development of a CS-PBM, because it yields utility values that compare better to the more frequently used generic instruments used in other economic evaluations, but only under the following circumstances: 1) there is no empirical evidence for insensitivity of the generic instrument, and 2) only use of mean utility values is intended rather than subgroup analysis (85) and 3) the health status or disease subtype of the sample on which the function was estimated is comparable to the sample on which the function is applied (111).

Findings here underline that the TTO health state values as modeled from a fractional factorial design can differ from direct TTO valuations of those states. Often but not always an OMEP is applied to allow the estimation of TTO values for all theoretically possible health states from only a fraction of health states. This study adopted that technique but also valued directly a selection of states that were observed frequently in patients. Using these states in the estimation of the preference algorithm resulted in lower scores for at least some of these states (Figure 4.1). These results suggest that discrepancies exist between modeled TTO values and directly observed TTO values for the most occurring health states which may affect the validity of the measure. Little guidance is available for researchers who wish to design a valuation study for a CS-PBM using state-of-the-art techniques, so it is not surprising that practices vary and this deserves more attention to

ensure that high quality CS-PBMs are produced. Ideally the process of constructing the CS-PBM is supported by the original developers of the questionnaires. This is relevant for example to avoid wild grow of value sets (e.g. for the QLQ-C30 now multiple value sets exists derived via mappings (90,100,111,117), to further guarantee quality, and to offer support to users of the CS-PBM .

Constructing and using a CS-PBM for the purpose of resource allocation could be considered when the following conditions are met: empirical evidence disproves sensitivity of existing generic instruments, empirical evidence proves the superiority of the condition-specific measure from which the new preference-based measure will be derived, and the derived CS-PBM is shown to be superior to the existing CS-PBM, not just in terms of statistical sensitivity, but also in terms of absolute differences. The development of CS-PBMs is welcome from an academic point of view as it pushes methodological frontiers and introduces new data for comparing measures in a field where no gold standard PBM exists. Use in resource allocation of these instruments, however, is only warranted when the above mentioned conditions are met. The introduction of preference-based measures which are specific to a certain disease has the presupposed merit of sensitivity to disease-specific effects of interventions, but this article shows that such an advantage is not necessarily achieved. Furthermore, the possible increase in sensitivity is traded off to the loss of comparability of absolute differences in utility values, which are most important for economic evaluations. It is argued here that without convincing empirical evidence on the insensitivity of a generic instrument, using a CS-PBM introduces confusion about the appropriate outcome measures in cost-utility analysis and health-care decision making.

ACKNOWLEDGEMENTS

We would like to thank for their Professor B. Uitdehaag from the multiple sclerosis centre of the VU Medical Centre in Amsterdam and Dr. J.Luime from the Netherlands Expert Centre for Work-Related Musculoskeletal Disorders, University Medical Center Rotterdam for sharing their expertise. Furthermore we would like to thank Mike Horton, from the University of Leeds for his suggestions for the Rasch analysis and Mark Oppe for sharing his thoughts on the design of the TTO study. We owe gratitude to Bart Groenendijk for his aid with setting up the computer assisted experiment, to Ming Au for automating the data extraction and to Fleur van de Wetering and Sandra de Vries for their assistance during the TTO exercise.

SUPPLEMENTARY MATERIAL

Results from the process of item selection

Item selection: HAQ

The 20 items of the HAQ did not have disordered thresholds. The HAQ had some misfit to the Rasch model, caused by item 10 ('wash body') and 16 ('open jar') with significant fit residuals (>3). Removal of these items improved the fit of the scale to the Rasch model (Item fit residual = -0.37, SD = 1.36, Item trait interaction χ^2=91, DF=90, p=0.44, Person separation index = 0.94). All remaining 18 items performed nicely in the Rasch model, and none of the remaining 18 items showed differential item functioning for gender or age group (consisting of two groups under and over the median age of 53).

Employing further psychometric criteria did not aid the selection of items as all HAQ items had >37% of the responses on the highest level. Linear correlation between individual HAQ items and EQ-5D index was <0.29, with an average of 0.21 (SD = 0.03) without any marked deviation.

On the basis of the analysis of three previous Rasch analyses (109,114,118) it was decided to go with the five items that are included in both the HAQ and its successor the HAQ-II. These items were 'Get on and off the toilet', 'Open car doors', 'stand up from a straight chair', 'walk outdoors on flat ground' and 'reach and get down a five pound object (such as a bag of sugar) from just above your head'. The last item (5 pound object), is different in the Dutch translation of the HAQ, where 5 pounds is changed in to 1 kg, meaning that it represents less disability than the original item (3), which may yield a smaller utility decrement in a TTO study for that item than when using the original English version.

Item selection: MSIS-29

The Rasch analysis was applied separately for the physical and psychological functioning scales of the MSIS-29.Nearly all items of the physical scale showed difficulty for respondents to differentiate between the categories 'a little', 'moderate' & 'quite a bit' and some had reversed thresholds (no ordinal order of categories). All items were rescored to a 4-point scale merging 'moderate' with one of the adjacent categories. Rescoring improved the fit considerably but the scale continued to misfit the Rasch model. Several items were deleted for different reasons (Table S4.1). Applying psychometric criteria suggested to the removal of item 17 'trouble using transport' for which 45% of respondents reported no problems. Item 15 was retained despite 29% of respondents reporting no problems on the item, as removal worsened scale fit. The resulting scale was found to be unidimensional, showed no DIF and fit the Rasch model. On the basis

of the spread of difficulty represented by the item, and advice from the clinical expert, 5 out of 8 items were selected for the vignette. The selected items are 4 ('Problems with your balance'), 6 ('Being clumsy'), 13 ('Limitations in your social and leisure activities at home'), 15 ('Difficulties using your hands in everyday tasks'), 16 ('Having to cut down the amount of time you spent on work or other daily activities').

The psychological scale of the MSIS-29 showed disordered thresholds for item 26. Items showed difficulty for respondents to distinguish between level 'moderate' and the two adjacent categories. All items were rescored to a 4-point scale merging either with the adjacent higher or lower level of level 'moderate'. Rescoring slightly improved model fit. Table S4.1 shows the process of deleting items and the effect on the total fit of the psychological scale. The resulting items fit the Rasch model individually and as a scale, showed no DIF and were unidimensional. On the basis of the psychometric criteria, item 22 ('problems sleeping') and 29 ('feeling depressed') were not considered to be a candidate for the selection on the vignette as 30% of respondents reported 'no problems'. Item 28 was not considered for the vignette on the basis of spread of difficulty of the item. The linear correlation (R^2) between the items (i) and sum of the other (not selected) items in the domain was used to inform a final decision. Linear correlation was highest (>.6) for item 25 ('feeling anxious or tense'), 26 ('feeling irritable, impatient or short tempered') and 27 ('problems concentrating'). Despite the high correlation for item 25, it was decided to go with items 26 and 27 as item 25 had a high fit residual (Fit. Res. -2.1, $\chi^2_{(df=5)}$= 9.7, p=.08). Upon consultation of the clinical MS expert it was decided to add item 23 'feeling mentally fatigued' to the final list, as this item was deemed a crucial element in MS.

Item selection: QLQ-C30

The QLQ-C30 questionnaire consists of five functional scales, nine symptom scales and one global health status. The aim of the item selection was to include only one item of each QLQ-C30 scale, and to have the health state represent all the dimensions identified by factor analysis. Rasch analysis was not performed on all of the scales, as some scales consist of only one item. Therefore, psychometric criteria combined with expert opinion were used as the main criteria for selection of the items.

A principal component analysis was performed on all items but global health status. The global health status scale is an assessment of the quality of life in general rather than a specific aspect of quality of life and was therefore not considered for the health state description. Five different factors were identified, which we summarize with the following factor identifiers: physical functioning, vitality, mental functioning, discomfort and pain (Table S4.2). The fourteen QLQ-C30 scales loaded on five factors. Within these

factors, we aimed to select items that belonged to different QLQ-C30 scales to obtain a maximum representation of relevant items that impact on quality of life. When items loaded on the same factor but belonged to different QLQ-C30 scales, we accepted a maximum correlation of 0.6 between the two items.

The 14 QLQ-C30 scales were rank ordered based on their linear correlation with the global health scale. An arbitrary cut off point was that scales had to explain 15% of the variance in the global health scale to be selected for the health state. Following this strategy the following QLQ-C30 scales were not considered for the vignette 'Nausea and vomiting', 'Dyspnoea', 'Constipation', 'Diarrhea' and 'Financial difficulties' Some of these scales consist of multiple items so items were also rank ordered on the percentage of variance explained on the global health scale and on their distribution of responses on the item levels. We based our final selection of items within a QLQ-C30 scale on this rank order. In one instance we deviated from our strategy. We chose to include 'have you felt nauseated' rather than 'appetite loss'. Although the item 'have you felt nauseated' did not meet our prior requirements, the item 'appetite loss' was replaced with 'have you felt nauseated' on the advice of the cancer expert.

SUPPLEMENTARY MATERIAL

Table S4.1 Steps in Rasch analysis for item selection

	Actions taken	Items deleted	Reason for deletion	Item-trait Chi-square (df)	Probability	Mean fit residual (sd)	Person separation index
HAQ	1 Full model	None		136 (df)	<0.00	-0.40 (1.59)	0.95
	2 Final model / Selection of items based on literature	All but item 3, 8, 12, 13,15	Items were not selected for new HAQ_II instrument	22 (25)	0.64	-0.20 (1.0)	0.83
MSIS-29	*Physical domain*						
	1 Full model	None		278 (100)	<0.00	-.03 (3.1)	0.95
	2 Rescore items to 4 point scale	None		283(100)	<0.00	-.08 (2.9)	0.95
	3 Remove items	5, 10, 9, 20, 7	Bonferroni significant misfit	108 (75)	0.008	-.29 (2.0)	0.94
	4 Remove items	8, 12, 18	Fit,res >2,5	73 (60)	0.019	-.14 (1.3)	0.93
	5 Remove items	17	45% reports 'no problems'	73 (60)	0.12	-.13 (1.3)	0.93
	6 Final model / remove items	3, 2, 1	Achieve uni-dimensionality	32 (40)	0.81	-.00 (1.0)	0.89
	Psychological domain						
	1 Full model	None		104 (45)	<0.00	.05 (2.6)	0.89
	2 Rescore items	None		84 (45)	0.0004	-.13 (2.2)	0.89
	3 Remove items	22	Bonferroni significan misfit	40 (40)	0.48	-.09 (1.7)	0.89
	4 Remove items	25	Fit,res >2,5	39 (40)	0.48	-.09 (1.7)	0.89
	5 Final model / remove items	24	DIF for agegroup	36 (35)	0.42	-.00 (1.5)	0.88
QLQ-C30	1 Full model	None		129 (56)	<0.00	-0.07(1.5)	0.9
	2.Final model	all but item 2, 6, 9, 14,18, 20, 22,27	Items did not met psychometric criteria or had disordered thresholds	31 (16)	0.01	0.07 (1.4)	0.82

Table S4.2 Factor structure derived from PCA

Physical functioning		Vitality		Mental functioning		Discomfort		Pain	
Taking a long walk (PF)	**0.72**	**Social activities (SF)**	**0.66**	Depressed (EF)	0.75	**Nauseated (NV)**	**0.80**	**Pain (PA)**	**0.84**
Strenuous activities (PF)	0.68	Family life (SF)	0.66	Tense (EF)	0.74	Vomited (NV)	0.77	Pain interfere with daily activities (PA)	0.81
Tired (FA)	**0.66**	Help with eating. dressing washing or using the toilet (PF)	0.66	Irritable (EF)	0.74	Appetite loss (AP)	0.62		
Short of breath (DY)	0.60	Stay in bed or chair (PF)	0.64	**Worry (EF)**	**0.73**	Constipation (CO)	0.46		
Limited in work or other daily activities (RF)	**0.60**	Short walk (PF)	0.59	**Concentrating on things (CF)**	**0.51**	Diarrhea (DI)	0.36		
Need to rest (FA)	0.59	Limited in hobbies (RF)	0.57	Sleeping (SL)	0.49				
Felt weak (FA)	0.57	Financial difficulties (FI)	0.47	Remembering things (CF)	0.43				

QLQ-C30 scale abbreviations: PF=physical functioning; RF=role functioning; EF=emotional functioning; CF=cognitive functioning; SF=social functioning; FA=fatigue; NV=Nausea and vomiting; PA=pain; DY=dyspnoea; SL=insomnia; AP=appetite loss; CO=constipation; DI=diarrhea; FI=financial difficulties.
Items in bold were selected for the QLQ-C30 health states.

Chapter 5

Impaired health-related quality of life in acute myeloid leukemia survivors: a single-center study

With W. Ken Redekop, Carin. A. Uyl-de Groot and Bob Löwenberg

Published in European Journal of Haematology 2014(93): 198-206

ABSTRACT

The purpose of this study was to assess the impact of acute myeloid leukemia (AML) and its treatment on health-related quality of life (HRQOL) by comparing the HRQOL of AML survivors with the HRQOL in the general population.

Two HRQOL questionnaires (EQ-5D and QLQ-C30) were sent to patients diagnosed with AML between 1999-2011 at a single academic hospital and still alive in 2012. HRQOL in AML survivors was compared with general population reference values. Multivariate analysis was used to identify factors associated with HRQOL in AML survivors.

Questionnaires were returned by 92 of the 103 patients (89%). AML survivors reported significantly worse functioning, more fatigue, pain, dyspnea, appetite loss and financial difficulties and lower EQ-VAS scores than the general population ($P<0.05$). Impaired HRQOL in AML survivors was mainly found in survivors without a paid job. Other factors associated with a poor HRQOL were allogeneic hematopoietic stem cell transplantation and the absence of social support.

This single-center study showed that the HRQOL in AML survivors is worse than the HRQOL in the general population. HRQOL in these patients can be improved by adequately treating and preventing fatigue, pain, dyspnea and appetite loss.

INTRODUCTION

Treatment outcome of acute myeloid leukemia (AML) has improved over the past decades. This improvement can be attributed to changes in chemotherapy regimens, the introduction of allogeneic and autologous hematopoietic stem cell transplantation (HSCT), better supportive care to prevent treatment-related mortality and risk-adapted treatment approaches (119-121). As a consequence of the increasing number of AML survivors, it is becoming more important to assess the impact of AML and its treatments on health-related quality of life (HRQOL).

Health-related quality of life is a broad concept which covers different domains such as physical, mental, social and role functioning (122). Information about the impact on HRQOL can be used for different purposes. First, this information is useful for treatment allocation. Currently, treatment allocation in AML depends solely on the effectiveness of the different treatments in terms of survival (50). Due to the aggressive nature of the available treatments, HRQOL should be included as additional criterion in treatment decisions. Furthermore, HRQOL information also provides insight into specific health problems and treatment needs of patients with AML. The identification of these health problems can help in the effort to improve current treatments and develop new treatment modalities (123,124).

Despite the importance of evidence-based information about HRQOL in patients with AML, only limited information of HRQOL in AML is currently available (53,125). A few studies have found that patients who received allogeneic HSCT had significantly worse HRQOL compared to those who received autologous HSCT or high-dose chemotherapy (54-56). Furthermore, it was found that HSCT recipients had poorer physical and mental health compared to the general population (126). However, only 22% of the patients in that study were treated for AML. Specific information regarding the impact of AML and its treatments on HRQOL is therefore lacking. The aim of our study was to compare the HRQOL of AML survivors with that in the general population. An additional aim was to gain further insight into the impact of AML on HRQOL by exploring factors associated with HRQOL in AML survivors.

PATIENTS AND METHODS

Patient selection

Questionnaires were sent in 2012 to all AML patients who participated in clinical trials HOVON-29, HOVON-42(A), HOVON-43, HOVON-81, HOVON-92(127-131) between 1999

and 2011 and were diagnosed and had received first-line treatment at the Erasmus University Medical Center (Erasmus MC) in Rotterdam (N=103). Patients were only included if they had still follow-up visits planned at the Erasmus MC. The main reasons for loss to follow-up were migration to other areas within or outside the Netherlands and scheduled follow-up visits in other hospitals closer to their homes.

Together with the questionnaires, patients received an invitation to participate by mail. Patients were informed that they gave permission to participate in the study by returning the questionnaire. The study protocol had been approved by the institutional medical ethics committee.

Sociodemographic and clinical data

Sociodemographic data, including age, sex, education, employment status, living arrangement and social support, were collected via the questionnaire. Social support was measured by asking whether the patient received social support from family and/or friends. Clinical data were obtained via the HOVON Data Center, including date of diagnosis, last treatment received and leukemia recurrence.

Quality of life measurement

HRQOL was measured with the EuroQol-5 Dimension (EQ-5D) and the European Organization for Research and Treatment of Cancer quality of life questionnaire (QLQ-C30). The EQ-5D is a generic quality of life questionnaire and consists of two parts: the EQ-5D descriptive system and the EQ - visual analogue scale (EQ-VAS). The descriptive system comprises five questions related to the dimensions 'mobility', 'self-care', 'usual activities', 'pain/discomfort' and 'anxiety/depression' (132). The most recent version of the EQ-5D contains five response levels for each question (133,134). The descriptive system was converted to a health utility index with 0 representing death and 1 perfect health (135). The EQ-VAS asks respondents to rate their current health state on a scale from 0 (worst imaginable health) to 100 (best imaginable health).

The QLQ-C30 is a cancer-specific questionnaire and consists of 30 questions representing five functional scales (physical, role, cognitive, emotional and social functioning), a global quality of life scale and nine symptom scales (fatigue, nausea & vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea and financial difficulties). All scale scores range between 0-100. A higher score on the functional scales and the global quality of life scale indicates a better quality of life, while a higher score on the symptom scales indicates more symptoms and thereby a poorer quality of life (89,136).

HRQOL in the general population

Published studies were used to estimate general population HRQOL for the patients in this study (137,138). The QLQ-C30 reference values were estimated by matching age and gender. The EQ-VAS reference values were calculated by using a regression adjustment including age, gender, years of education, paid employment, income and living arrangement. Education and income were not measured in the AML population. In order to estimate general population values, it was assumed that the distribution of education and income was similar to the distribution in the general population. General population values for the EQ-5D with five answer levels were not yet available.

Exploration of factors associated with HRQOL in patients with AML

Several factors that are possible associated with HRQOL in patients with AML have been identified based upon theoretical HRQOL models found in the literature (139,140). The following characteristics were included as predictor variables in the analyses: current age, gender, education level, living arrangement, leukemia recurrence, type of last treatment received, time since diagnosis, employment status and social support. These characteristics are either personal, environmental or biological factors which influence quality of life according to Wilson & Cleary (139) or personal or social support factors identified by Holland (140).

Statistical analysis

Descriptive statistics were generated to describe the sociodemographic and clinical characteristics and the HRQOL scores of the study sample. The two sided (paired) t-test and chi-square test were used to test for differences in patient characteristics and HRQOL scores between the AML survivors and the general population (significance criterion: $P<0.05$). Significant differences in QLQ-C30 scores ≥ 10 points were considered clinically relevant (141). Univariate analyses, using Mann-Whitney U test, Kruskal-Wallis test, Spearman rank correlation test, t-test or ANOVA, were used to assess whether subgroups with a poorer HRQOL could be identified.

Multivariate linear regression analyses using backward selection were performed to identify patient and clinical characteristics associated with the HRQOL in patients with. Correlation coefficients between all independent variables were evaluated to test for multicollinearity. No strong correlations existed between the variables, with exception of the correlation between age and retirement. These two variables were combined in the multivariate analyses. Three different groups were identified: age <=55 years, age >55 years and retired, age >55 years and not retired. Due to the small sample size and the exploratory nature of this study, only variables with a P-value > 0.1 were excluded.

Separate analyses were performed for the EQ-5D utility score, the EQ-VAS score and the different scales of the QLQ-C30.

A bootstrap sensitivity analysis was performed to assess the robustness of the results of the backward selection procedure (142). The backward selection procedure was repeated in 1000 bootstrap samples with 92 patients each (i.e., the number of patients in the original study population). These samples were randomly drawn with replacement from the original sample. We identified the total number of unique models in the bootstrap analysis as well as the most frequently selected model. Furthermore, we determined how often individual variables were included in the selected models.

RESULTS

Patient characteristics

In total, 92 patients with AML (89%) returned the HRQOL questionnaire. The sociodemographic and clinical characteristics of the patients are shown in Table 5.1. The mean age in the AML sample was 52.7 years (median (range): 55(25-78) years). The age was comparable with the mean age in the QLQ-C30 reference population, but significantly higher than the mean age in the EQ-VAS reference population. No significant differences were found in living arrangement and gender between the AML sample and reference populations. The percentage of retired patients or patients with a paid job was significantly lower in the AML study sample compared to the QLQ-C30 reference population. A large proportion of the AML patients without a paid job were unable to work due to the disease or other health problems. About 60% of the patients had received an allogeneic HSCT, either as part of the first line of treatment or following leukemia recurrence. In total, 18% of the patients in our cohort had experienced a leukemia recurrence.

Generic health-related quality of life (EQ-5D)

Only a minority of the patients reported problems with self-care (9%) and about a quarter of the patients (27%) reported anxiety. Relatively speaking, patients more frequently reported pain and problems with usual activities or mobility. Less than 10% of the patients reported severe or extreme problems on any of the health dimensions (Figure 5.1).

The EQ-5D utility of all patients with AML in this study ranged from 0.21 to 1.0, with a mean of 0.82 (Table 5.2). The average EQ-VAS score of the patients was significantly lower than the predicted general population EQ-VAS (74.6 and 78.8, respectively). Patients aged < 65 years without a paid job had a significantly lower EQ-5D utility and EQ-VAS score compared to other patients (Table 5.3). The lower utility score was related

Table 5.1 Patient characteristics

	AML patients N=92		Healthy controls(137,138) (QLQ-C30) N=1,731		(EQ-5D) N=2,367	
Age (years)						
Mean (SD)	52.7 (12.8)		52.9 (15.7)		48.4 (16.4)***	
Median (Range)	55 (25-78)					
Time since diagnosis (years)						
Mean (SD)	5.9 (2.7)					
Median (Range)	6 (2-13)					
Time since last treatment (years)						
Mean (SD)	5.3 (2.8)					
Median (Range)	5 (1-12)					
	N	%	N	%	N	%
Sex						
Female	45	49	796	46	1,337	51
Male	47	51	935	54	1,030	49
Level of education						
Elementary school or secondary education	28	30	774	44.7		
Vocational school	37	40	318	18.4		
University	26	28	635	36.7		
Unknown	1	1	4	0.2		
Employment status						
Paid job	40	43	880	50.8**		
Retired	19	21	439	25.4**		
Other	32	35	412	23.8**		
Unknown	1	1				
Ethnicity						
European	81	88				
Other	9	10				
Unknown	2	2				
Living arrangement						
Living with a partner	72	78	1,333	77	1,581	71**
Living without a partner	20	22	373	22	786	29**
Unknown			25	1		
Presence of social support						
Yes	87	96				
No	4	4				
Last treatment received						
Chemotherapy	29	32				
Autologous HSCT	9	10				
Allogeneic HSCT	54	59				
Relapse (%)						
Yes	17	18				
No	75	82				

** Statistical significant at α = 0.05

*** Statistical significant at α = 0.01

Table 5.2 QLQ-C30 and EQ-5D mean scores in AML survivors and the general population

	n	% reporting problems	Mean score (SD) AML patients		Predicted score general population*	Difference in score	P-value
QLQ-C30							
Physical functioning	91	73	80.4	(18.7)	91.3	-10.9 [a]	<0.0001
Role functioning	92	55	74.8	(26.9)	89.3	-14.5 [a]	<0.0001
Emotional functioning	90	60	83.1	(20.2)	89.0	-5.9	0.0075
Cognitive functioning	90	61	77.6	(24.1)	92.7	-15.1 [a]	<0.0001
Social functioning	89	57	81.3	(23.8)	94.5	-13.2 [a]	<0.0001
Global quality of life	91	82	75.3	(19.5)	78.0	-2.7	0.1922
Fatigue	91	78	35.3	(27.6)	17.1	18.2 [a]	<0.0001
Nausea/Vomiting	91	19	6.2	(18.4)	2.7	3.5	0.0688
Pain	91	47	21.8	(28.9)	14.8	7.0	0.0248
Dyspnoea	91	36	18.3	(27.3)	6.9	11.4 [a]	0.0002
Insomnia	90	39	19.6	(28.2)	13.8	5.8	0.0521
Appetite loss	91	21	9.2	(19.9)	3.2	6.0	0.0054
Constipation	90	12	5.2	(14.9)	4.5	0.7	0.6835
Diarrhoea	91	11	5.5	(18.1)	3.9	1.6	0.4167
Financial difficulties	90	33	19.2	(31.2)	3.1	16.1 [a]	<0.0001
EQ-5D							
EQ5D utility score	88	-	0.82	(0.17)			
EQ VAS score	91	-	74.6	(17.4)	78.8	-4.2	0.0333

* The scores are matched according to relevant patient characteristics. These characteristics were sex and 10-year age group for the QLQ-C30 and age (centered at median), sex, living arrangement, employment, education and household income for the EQ-VAS score. As years of education and income level were unknown in the AML population, a weighted average was calculated for the EQ-VAS assuming a similar distribution as in the general Dutch population, a Indicates a clinically relevant change (>10 points)

to more problems with mobility and usual activities and more anxiety/depression. Allogeneic HSCT, younger age and absence of social support were also associated with a lower EQ-VAS score.

Cancer-specific quality of life (QLQ-C30)

The majority of the patients with AML reported problems on the five functioning scales of the QLQ-C30 (Table 5.2). The average scores on all functioning scales were significantly lower in patients with AML compared to adjusted general population scores. The differences in physical, role, cognitive and social functioning were also clinically relevant. Despite these differences, no significant difference was found for the global quality of life.

Table 5.3 Subgroup analysis for the three global quality of life measures

	QLQ-C30 - QL		EQ5D utility		EQ-VAS	
Age						
<=55 years	72	P=0.02	0.81	P=0.39	72	P=0.01
> 55 years	78		0.83		78	
Gender						
Female	74	P=0.57	0.79	P=0.15	73	P=0.46
Male	76		0.85		76	
Education level						
Primary or secondary education	74	P=0.67	0.79	P=0.39	75	P=0.82
Vocational school	76		0.83		75	
University	76		0.86		74	
Living arrangement						
Living without a partner	77	P=0.31	0.83	P=0.22	75	P=0.49
Living with a partner	71		0.77		72	
Employment status						
Paid job	80	P<0.01	0.88	P<0.01	80	P<0.01
Retired	81		0.81		79	
Age < 65 years, no paid job	65		0.75		65	
Social support						
No social support	48	P=0.07	0.62	P=0.11	53	P=0.09
Social support	77		0.83		76	
Time since diagnosis						
<= 5 years	74	P=0.85	0.81	P=0.99	73	P=0.65
> 5 years	76		0.83		77	
Last treatment received						
High-dose chemotherapy/Autologous HSCT	81	P=0.02	0.83	P=0.77	78	P=0.08
Allogeneic HSCT	71		0.82		72	
Relapse						
No relapse	76	P=0.65	0.83	P=0.19	75	P=0.57
Relapse	72		0.78		73	

QL = global quality of life scale, HSCT = hematopoietic stem cell transplantation

Fatigue was the most frequently reported symptom in patients with AML (78%). Other frequently reported symptoms were pain, dyspnea, insomnia and financial difficulties. Patients with AML had significantly more problems with fatigue, pain, dyspnea and appetite loss than the general population. Furthermore, financial difficulties were more frequently reported by patients with AML. Only the differences in fatigue, dyspnea and financial difficulties were clinically relevant.

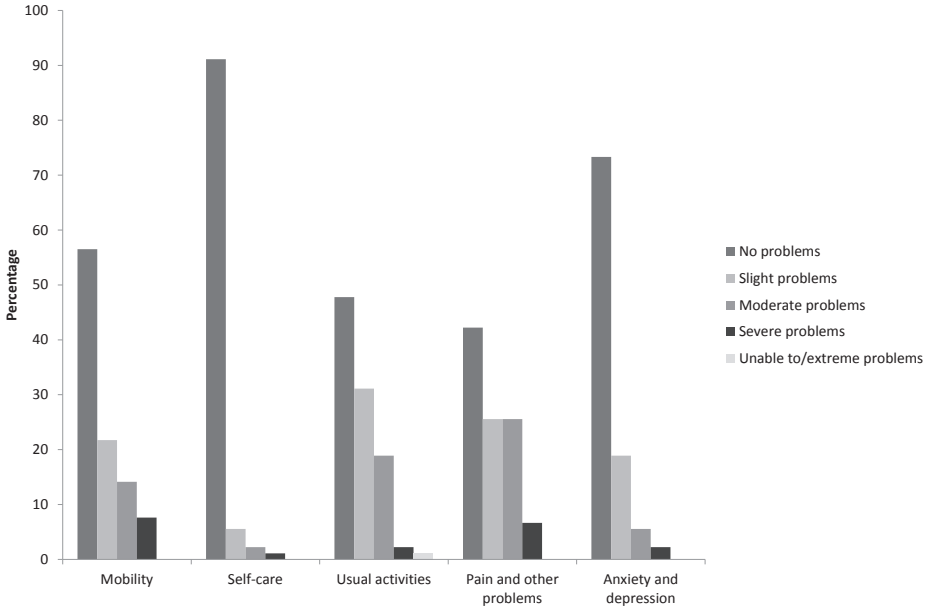


Figure 5.1 Reported problems on the EQ-5D domain

Factors associated with quality of life in AML patients

The multivariate analysis showed that about 20-30% of the variance in scores on overall HRQOL could be explained by differences in patient and disease characteristics, including age, employment status, social support and last treatment received (Table 5.4). The absence of a paid job was associated with a poorer HRQOL in almost all scales. Older, non-retired patients had better physical and role functioning scores and reported fewer problems with fatigue and dyspnea. Furthermore, a better overall HRQOL was indicated for these patients. The availability of social support was associated with better physical, role and social functioning, fewer symptoms of dyspnea, less financial difficulties and better overall HRQOL. Patients who had received an allogeneic HSCT reported worse physical and social functioning, more problems with fatigue and dyspnea and a significantly lower overall HRQOL. Leukemia recurrence was not significantly associated with functioning or overall quality of life. However, patients with recurrence of leukemia reported fewer symptoms of diarrhea. A longer time since diagnosis was significantly associated with better social functioning and fewer problems with diarrhea, nausea and vomiting.

The bootstrap analysis resulted in 200-350 unique models selected by backward selection in the 1000 bootstrap samples. The models from the original patient sample (see Table 5.4) were selected in 2-16% of the bootstrap samples. Inconsistencies were mainly found in variables with a P-value close to 0.1 in the original model. Most variables in-

Table 5.4 Multivariate analysis of factors associated with health-related quality of life in AML survivors

		QLQ-C30															EQ5D	
		PF N=88	RF N=89	SF N=86	EF N=87	CF N=87	QL N=88	FA N=88	NV N=88	PA N=88	DY N=88	SL N=87	AP N=88	CO N=87	DI N=88	FI N=87	EQ5D utility N=85	VAS score N=88
R2		0.30	0.27	0.31	0.15	0.15	0.31	0.29	0.04	0.07	0.31	0.10	0.05	0.06	0.14	0.09	0.19	0.30
Intercept	Mean	66.1***	37.1**	20.9***	89.4***	84.8***	53.0***	37.3***	14.2***	16.7***	61.9***	13.2***	5.8**	2.3*	15.9***	50.0***	0.72***	52.0***
	SE	(9.3)	(15.2)	(12.3)	(2.4)	(2.8)	(10.5)	(7.9)	(4.7)	(3.7)	(14.6)	(3.4)	(2.6)	(1.9)	(4.9)	(15.1)	(0.08)	(8.9)
Age > 55 and not retired (vs age <= 55 years)	Mean	10.5**	21.0***	NS	NS	NS	9.4**	-11.0*	NS	NS	-13.2**	NS	NS	NS	NS	NS	NS	10.0**
	SE	(0.2)	(6.2)				(4.3)	(6.1)			(6.1)							(3.8)
Age > 55 and retired (vs age <= 55 years)	Mean	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	-15.5*	-0.08*	NS
	SE															(8.0)	(0.04)	
Female	Mean	NS	9.4*	NS	NS	NS	7.3*	-9.4*	NS	NS	-18.6***	NS	NS	NS	NS	NS	NS	6.6*
	SE		(5.6)				(4.0)	(5.6)			(5.6)							(3.5)
Low level education	Mean	NS	NS	NS	NS	NS	NS	NS	NS	NS	15.6***	NS	NS	NS	7.1*	NS	NS	NS
	SE										(5.7)				(4.0)			
Living together with a partner	Mean	NS	10.9*	19.9***	NS	NS	NS	-12.0*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
	SE		(6.5)	(5.7)				(6.8)										
Relapse	Mean	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	-8.9*	NS	NS	NS
	SE														(4.7)			
Allogeneic HSCT	Mean	-6.2*	NS	-8.1*	NS	NS	-8.4**	9.5*	NS	NS	12.1**	NS	NS	NS	NS	NS	NS	NS
	SE	(3.5)		(4.4)			(3.7)	(5.2)			(5.1)							
Time since diagnosis	Mean	NS	NS	1.6*	NS	NS	NS	NS	-1.3*	NS	NS	NS	NS	NS	-1.8*	NS	NS	NS
	SE			(0.8)					(0.7)						(0.7)			
No paid job	Mean	-16.9***	-21.7***	NS	-15.8 ***	-18.2 ***	-17.5***	27.8***	NS	16.1**	10.8*	17.8***	9.2**	7.7**	NS	NS	-0.12***	-18.1***
	SE	(3.8)	(5.8)		(4.0)	(4.8)	(4.1)	(5.9)		(6.4)	(5.8)	(5.9)	(4.4)	(3.3)			(0.04)	(3.6)
Support family and friends	Mean	21.7**	27.8**	41.6***	NS	NS	28.6***	NS	NS	NS	-49.0***	NS	NS	NS	NS	-29.0*	0.18**	23.9***
	SE	(8.6)	(12.9)	(10.3)			(9.2)				(12.9)					(15.5)	(0.08)	(8.2)

* Significant at α = 0.1
** Significant at α = 0.05
*** Significant at α = 0.01
NS = Not significant
Factors selected by backwards selection using alpha <0.1.
PF = Physical functioning, RF = Role functioning, SF = Social functioning, EF = Emotional functioning, CF = Cognitive functioning, QL = Global quality of life, FA = Fatigue, NV = Nausea & Vomiting, PA = Pain, DY = Dyspnoea, SL = Insomnia, AP = Appetite loss, CO = constipation, DI = diarrhea, FI = Financial difficulties, HSCT = hematopoietic stem-cell transplantation.

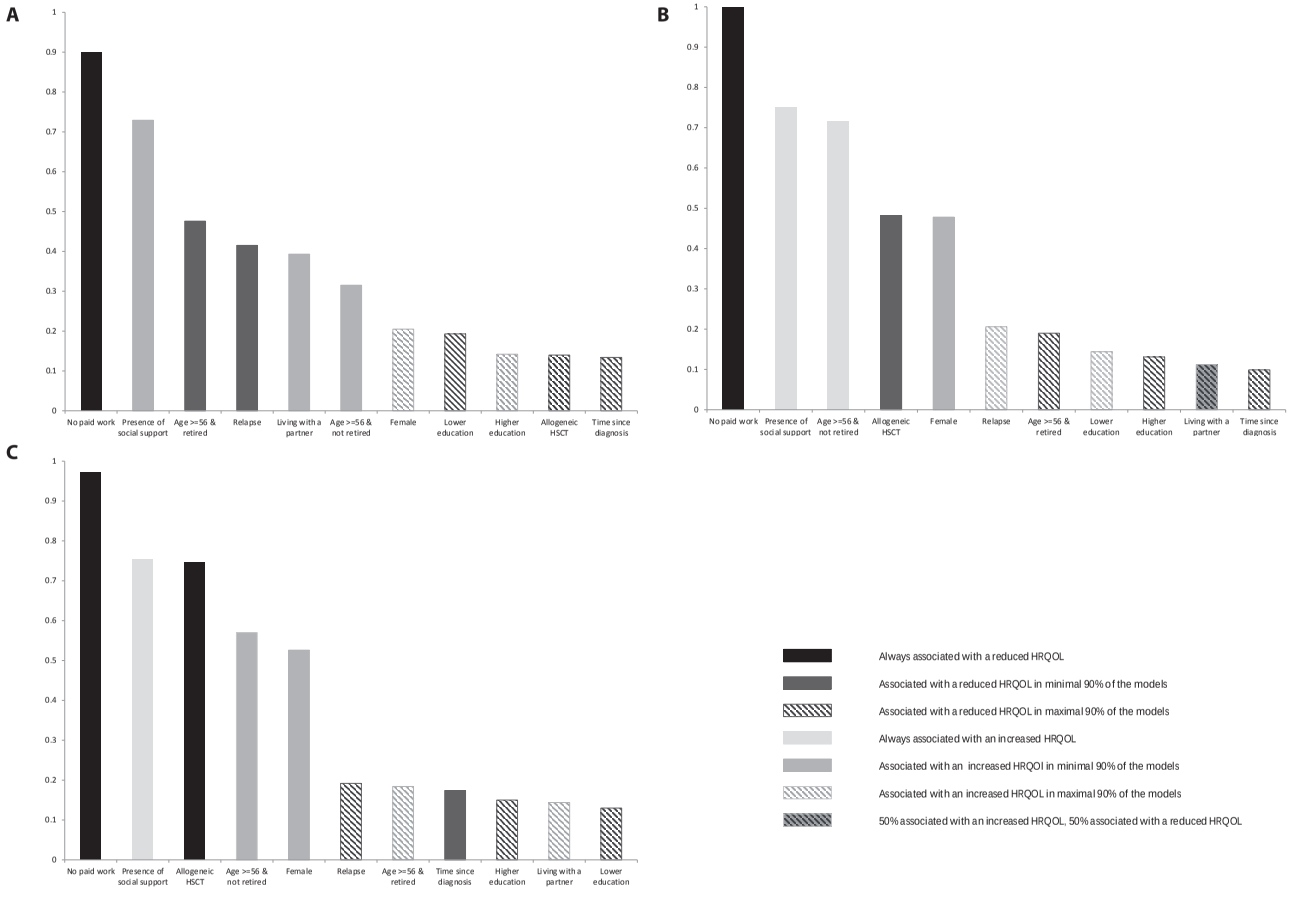


Figure 5.2 Frequency of variable selection in the bootstrap analysisfor A. the EQ-5D utility, B. the EQ-VAS and C. the Global quality of life scale of the QLQ-C30

cluded in the original model were selected in at least 50% of the bootstrap samples. Results for the three overall HRQOL scales are shown in Figure 5.2. The variable 'no paid job' was the most frequently selected variable for all three scales. The direction of the coefficients was not always consistent in the selected models. The inconsistency was larger for variables which were less often selected.

DISCUSSION

This study is the first one to compare HRQOL in AML survivors with the HRQOL in the general population. We found that AML survivors reported more severe problems related to fatigue, pain, dyspnea and appetite loss. Furthermore, they reported more financial difficulties and a lower level of functioning than the general population.

The impaired HRQOL was especially found in patients without a paid job. The majority of these patients is unable to work due to health problems. This is an important problem as the group accounts for about 25% of all AML survivors in our study. Clinicians should try to identify and treat the underlying health problems to avoid problems with returning to work. HRQOL questionnaires such as the QLQ-C30 can be used as diagnostic tool to identify the existing problems.

This study showed that fatigue is one of the most frequently reported health problems in AML survivors. Fatigue is also an important problem in other cancer survivors (143,144). Recent systematic reviews and meta-analyses showed that exercise programs are able to reduce fatigue for cancer survivors and patients treated with HSCT (145,146). However, further research is required to identify which type of physical activity program is most effective to reduce fatigue in AML survivors.

This study also provides some insights into factors associated with HRQOL in AML. The following factors were associated with a poor HRQOL in our study: lack of social support, allogeneic HSCT as post-remission treatment and younger age. However, due to the insufficient power of this study no definite conclusions can be drawn. Nevertheless, some of our findings might be reasonable based upon other available evidence.

The relationship between social support and quality of life has frequently been observed in other cancer types.(147-149) Patients with social support might be better able to cope with the disease, and consequently, have fewer problems with performing social activities than patients without social support. Consideration about how to improve social support is therefore worthwhile.

A poorer HRQOL after allogeneic HSCT is also found by other studies (54-56). The impaired HRQOL is mainly caused by graft-versus-host-disease (150). An allogeneic HSCT is often the preferred treatment option because it has the highest anti-leukemic effect and thereby the highest cure rate (65). However, a meta-analysis showed that an allogeneic HSCT has no survival benefit for all patients (151). As several studies have shown that the HRQOL is also impaired after allogeneic HSCT, it is recommended to combine both mortality and HRQOL in a comparison of post-remission treatments. Further research in larger patient samples is required for more evidence about the impact of allogeneic HSCT on both mortality and HRQOL in different AML subgroups.

The association between younger age and poorer HRQOL is remarkable, because other studies have found a better HRQOL in younger AML survivors (55,56). An important difference between these studies and our study is that we also included patients aged >60 years at time of diagnosis. These patients might use different internal values in the assessment of their HRQOL than younger patients (152-154). Younger patients are expected to have a paid job and take care of their children, while older patients have fewer responsibilities. Consequently, the same health problems might make it more difficult for younger patients to fulfill their roles at home and at work.

Although expected at the start of this study, no significant association has been found between leukemia recurrence and poorer HRQOL. The absence of a significant association might be related to the low number of survivors with a leukemia recurrence. The smaller the group of interest, the stronger the association needs to be in order to find a significant association. Furthermore, it is assumed that most patients with a leukemia recurrence have died and are therefore not included in this study. It might be possible that the patients who died had a poorer HRQOL than those patients who survived. However, longitudinal studies are needed to confirm these assumptions.

This study has several limitations. First, the AML survivors in the single academic hospital might not be representative for all AML survivors. A relatively large proportion of the survivors in our sample received an allogeneic HSCT. It might be possible that these patients are less often referred to local general hospitals for follow-up than other patients with AML. However, the high response rate (89%) in this study supports the validity of our results within the single academic hospital. Further research in a larger study setting is required to generalize these results to a broader population.

Second, no adjustment for multiple testing was performed and a relatively high significance level (0.1) was used to identify any possible associations between patient characteristics and HRQOL in our relatively small study sample. Consequently, some of

the associations found in this study are simply chance findings and our results need to be interpreted cautiously. This study should be viewed as a first exploration of factors associated with HRQOL in patients with AML and our findings should be replicated in other populations.

Finally, social support was not measured with a validated questionnaire including specific questions of the type of support. Instead, we asked generic questions about support from family and/or friends. These questions can be differently interpreted by the respondents which might bias the measurement of social support and its relation with HRQOL. For future studies, it is recommended to measure social support with a validated questionnaire, like the Social Provisions Scale, Inventory of Socially Supported Behaviors or ENRICHD Social Support Inventory (155).

This single-center study showed that the HRQOL in AML survivors is worse than the HRQOL in the general population. As the number of AML survivors will increase due to the improved survival rate, it becomes more and more important that new treatments not only focus on improvements in survival but also on improvements in HRQOL. According to the results of this study, most improvements are needed for the symptoms fatigue, dyspnea, pain, insomnia and for all functioning domains.

ACKNOWLEDGEMENT

The authors want to thank Bronno van der Holt from the HOVON Data Center, Erasmus MC for providing clinical data of the patients who participated in this quality of life study. Furthermore, the authors want to thank Sarah Lonergan from the Erasmus MC for her help regarding the logistics of this study and Marjolein Schouten for her help with the data management.

Chapter 6

How to measure quality of life utilities in acute leukemia patients? A comparison of the feasibility, validity and reliability of the generic questionnaire EQ-5D-5L and the disease-specific QLQ-PBM.

With W. Ken Redekop and Carin A. Uyl-de Groot

Submitted

ABSTRACT

The aim of this study was to assess the feasibility, validity and reliability of two recently developed preference-based instruments: the generic EQ-5D-5L and the cancer-specific QLQ-PBM, which is derived from the EORTC QLQ-C30, in patients with acute leukemia.

Questionnaires containing the EQ-5D-5L and QLQ-C30 were sent to patients who participated in HOVON clinical trials between 1999 and 2011 at a single academic hospital and were still alive in 2012. Feasibility was assessed according to the frequency of incomplete data. Validity was assessed by floor and ceiling effects, correlations with other quality of life scales and ability to distinguish between patients with different health status. Reliability was assessed by using the Cronbach's alpha.

Questionnaires were returned by 89% (111/125) of the patients. Only six and seven respondents did not fully complete the EQ-5D-5L and QLQ-PBM, respectively. The ceiling effect was much larger for the EQ-5D-5L than for the QLQ-PBM (31% versus 15%). Both questionnaires showed a good content validity according to strong correlations with other quality of life scales and both questionnaires were able to detect significant differences between patients with different health status.

This study showed evidence for the feasibility, reliability and discriminative ability of both preference-based questionnaires. The relatively high ceiling effect of the EQ-5D-5L is not considered a threat for the validity of the EQ-5D-5L in acute leukemia because the questionnaire was able to discriminate between relatively mild severity groups. However, our findings need to be confirmed in larger, longitudinal studies.

INTRODUCTION

Many different questionnaires are currently available to estimate health-related quality of life for use in economic evaluations. These questionnaires include generic questionnaires like the EQ-5D, the SF-6D and HUI3, as well as disease-specific preference-based measures (33). Furthermore, many algorithms have been developed to estimate generic preference-based utilities from disease-specific questionnaires (32). It has been shown that the utility values of the questionnaires differ due to differences in content of the questionnaire, valuation technique and study population (33,107,156-160). These differences limit the comparability of economic evaluations if different questionnaires are used to calculate quality of life utilities. Therefore, the National Institute of Health and Care Excellence (NICE) in the UK has explicitly stated that the EQ-5D should be used to measure quality of life utilities, unless there is proof that the EQ-5D is not valid in the target population (161).

Evidence about the validity of the EQ-5D in acute leukemia patients is not yet available. Studies in other cancer types showed that the EQ-5D is a valid and reliable instrument to measure quality of life in cancer patients, although some concerns about content validity were reported (134,156,159,162-167). These concerns include the high percentage of patients reporting full health (ceiling effect) and the difficulty in detecting small changes in health. Since these shortcomings are also observed in other disease areas, the EuroQol group has recently developed the EQ-5D-5L, which has 5 instead of 3 levels (133).

Several studies have assessed the psychometric performance of the EQ-5D-5L by direct comparison with the 3 level instrument (EQ-5D-3L). These studies showed a significant reduction in ceiling effect for the EQ-5D-5L in both the general population and several disease populations, including cancer (168-170). Furthermore, it was shown that the new levels were useful additions to the instrument as the use of the new levels improved the discriminative ability of the instrument (168,169). Although these studies indicated an improved validity of the EQ-5D, information is needed regarding the validity of the EQ-5D-5L in comparison with cancer-specific questionnaires to adequately judge the validity of the EQ-5D-5L in specific cancer populations. At this moment, evidence is only available from one study which found that the EQ-5D-5L was non-inferior to the Functional Assessment of Cancer Therapy-Breast (FACT-B) questionnaire in Korean breast cancer patients (171). Further studies are needed to confirm these findings in other cancer populations. Therefore, the primary aim of this study was to assess the feasibility, validity and reliability of the EQ-5D-5L in acute leukemia patients in comparison with the cancer-specific EORTC QLQ-C30 questionnaire. A secondary aim of this study is to

assess the validity of a cancer-specific preference-based instrument, the recently developed QLQ-PBM (158). This instrument could be used to measure health-related quality if EQ-5D-5L proved to be insufficiently valid in acute leukemia patients.

METHODS

Patient selection

This study was a secondary analysis of a cross-sectional quality of life study (172). Questionnaires were sent to all acute leukemia patients who participated in clinical trials HOVON-29, HOVON-37, HOVON-42(A), HOVON-43, HOVON-70, HOVON-71, HOVON-81, HOVON-92 (127-130,173) between 1999 and 2011 and received first-line treatment at Erasmus University Medical Center (Erasmus MC) in Rotterdam, the Netherlands (N=125). Patients were excluded if no follow-up visits were planned in 2012 at Erasmus MC. The main reasons for loss to follow-up were migration away from the region (within or outside the Netherlands) and scheduled follow-up visits in other hospitals closer to their homes. Patients were informed that they would give permission to participate in the study by returning the questionnaire. The study was approved by the institutional medical ethics committee.

Sociodemographic and clinical data

The questionnaire included questions about the age, sex, education, employment status and marital status of the patient. Clinical data, including date of diagnosis, last treatment received and leukemia recurrence, were obtained via the HOVON Data Center.

Quality of life questionnaires and utility calculation

The EQ-5D-5L consists of the EQ-5D descriptive system and the EQ - visual analogue scale (EQ-VAS). The descriptive system includes five questions representing the dimensions mobility, self-care, usual activities, pain/discomfort and anxiety/depression (132). All questions have a 5-level response scale (133,134). The utility values were derived from an interim value set based upon a cross-walk of the EQ-5D-5L onto the EQ-5D-3L (135), because the results of direct valuation studies were not yet available. The EQ-VAS is a rating scale ranging from 0 (worst imaginable health) to 100 (best imaginable health).

The QLQ-PBM is a shortened version of the EORTC QLQ-C30 version 2. The QLQ-C30 is a cancer-specific questionnaire and consists of 30 questions (89,136). The QLQ-PBM only includes 8 items (trouble taking a long walk, limited daily activities, pain, nausea, fatigue, difficulty in concentrating, worrying and interference with social activities). These items were selected according to criteria of Rasch analysis, psychometric analysis and expert

opinion. The QLQ-PBM health states were valued with the time-trade-off method by the general public (158). Since a newer version of the QLQ-C30 was included in the current study, the question about trouble taking a long walk was recoded to a binary (yes/no) question.

Statistical analysis

The analysis of the psychometric properties of the two questionnaires included an analysis of the feasibility, validity and reliability of the questionnaires. The percentage of respondents with incomplete data was used as an indicator for the feasibility of the questionnaires. Incomplete data was defined as missing or inconclusive scoring of one or more items. The reliability was assessed using the Cronbach's alpha to assess the internal consistency of the scale.

The validity of the two questionnaires was evaluated by the floor and ceiling effects, the correlation with other measures of quality of life and the ability to distinguish between patients with different health status. The floor and ceiling effects were defined as the percentage of patients reporting the worst and best score, respectively.

The other measures of quality of life included two general quality of life scales: the EQ-VAS and the global quality of life scale of the QLQ-C30 (QL scale). Pearson or Spearman correlation coefficients were estimated between the dimension and utility scores of the two preference-based questionnaires and the general quality of life scales. We expected strong correlations between the preference-based questionnaires and the general quality of life scales ($r>0.5$) (174). Furthermore, Spearman correlation coefficients were estimated between the domains of the EQ-5D-5L and the QLQ-PBM to evaluate differences in content between the questionnaires.

Three different measures were used to distinguish patients with different health status. Both the EQ-VAS and QL scale were used as proxies for health status. Quartile scores were used to categorize these variables into four subgroups. Furthermore, patients aged <65 years were distinguished according to the self-reported ability to work. T-tests and ANOVA were used to test for significant differences in utility scores between the subgroups. Standardized effect sizes (ES) (Cohen's d) were calculated by dividing the difference in quality of life utility of subsequent categories by the overall standard deviation of the two categories. A small ES was defined as $d<=0.2$, a moderate ES as $0.2<d<0.5$ and a large ES as $d>0.8$ (174). All analyses were performed in SAS 9.2.

RESULTS

Patient characteristics

A total of 111 respondents returned the questionnaire (89%). The characteristics of the respondents are shown in Table 6.1. The respondents aged between 23 and 78 years with

Table 6.1 Patient characteristics

	N	%
Ethnicity		
European	97	87
Other	12	11
Unknown	2	2
Gender		
Female	53	48
Male	58	52
Level of education		
Elementary school or secondary education	32	29
Vocational school	49	44
University	29	26
Unknown	1	1
Employment status		
Paid job	48	43
Retired	21	19
Other	41	37
Unknown	1	1
Marital status		
Living with a partner	84	76
Living without a partner	27	24
Last treatment received		
Chemotherapy	34	31
Autologous HSCT	13	12
Allogeneic HSCT	64	58
Relapse		
Yes	91	18
No	20	82
Age (years)		
Mean (SD)	51 (13.4)	
Median (Range)	52 (23-78)	
Time since diagnosis (years)		
Mean (SD)	6 (2.7)	
Median (Range)	6 (2-13)	

SD = standard deviation, HSCT = hematopoietic stem cell transplantation

a median of 52 years. The time since diagnosis ranged between 2-13 years with a median of 6 years. About half of the respondents were male. The majority of the respondents had a European nationality and lived with a partner. About 60% of the respondents had a paid job or were retired; the remainder mainly consisted of patients who indicated to be unable to work due to cancer.

Feasibility of the questionnaires

Incomplete data was found in six and seven respondents for the EQ-5D-5L and QLQ-PBM, respectively. Only one respondent had incomplete data for both questionnaires. None of the respondents had incomplete data for all items of the questionnaire.

Validity

A fairly large proportion of patients reported no problems on the individual items of the EQ-5D-5L and the QLQ-PBM (Figure 6.1). The largest ceiling effects are found for the 'self-care' item of the EQ-5D-5L and the 'nausea' item of the QLQ-PBM. Overall, 31% of the respondents reported full health on the EQ-5D-5L compared to 14% for the QLQ-PBM. Of the patients in full health according to the EQ-5D-5L, 50-91% reported less than perfect health on the other quality of life scales (Table 6.2); the actual percentage depended on the scale used. These patients most frequently reported fatigue and problems with physical functioning on the QLQ-PBM. None of the patients reported the worst imaginable health according to the two instruments, but some patients reported the worst

Table 6.2 Quality of life for patients with perfect health according to the EQ-5D-5L (N=31)

	No perfect health		Mean (SD)	Median (Range)
	N	%		
QLQ-PBM utility score	17	(57)	0.94 (0.06)	0.93 (0.84-1.00)
EQ-VAS	29	(91)	86.0 (13.0)	90.0 (35.0-100)
Global quality of life scale	16	(50)	91.4 (10.5)	95.8 (58.3-100)
QLQ-PBM domains				
Physical functioning	10	(31)		
Role functioning	2	(6)		
Pain	1	(3)		
Nausea	3	(10)		
Fatigue	13	(42)		
Cognitive functioning	7	(23)		
Emotional functioning	5	(16)		
Social functioning	6	(19)		

SD = standard deviation

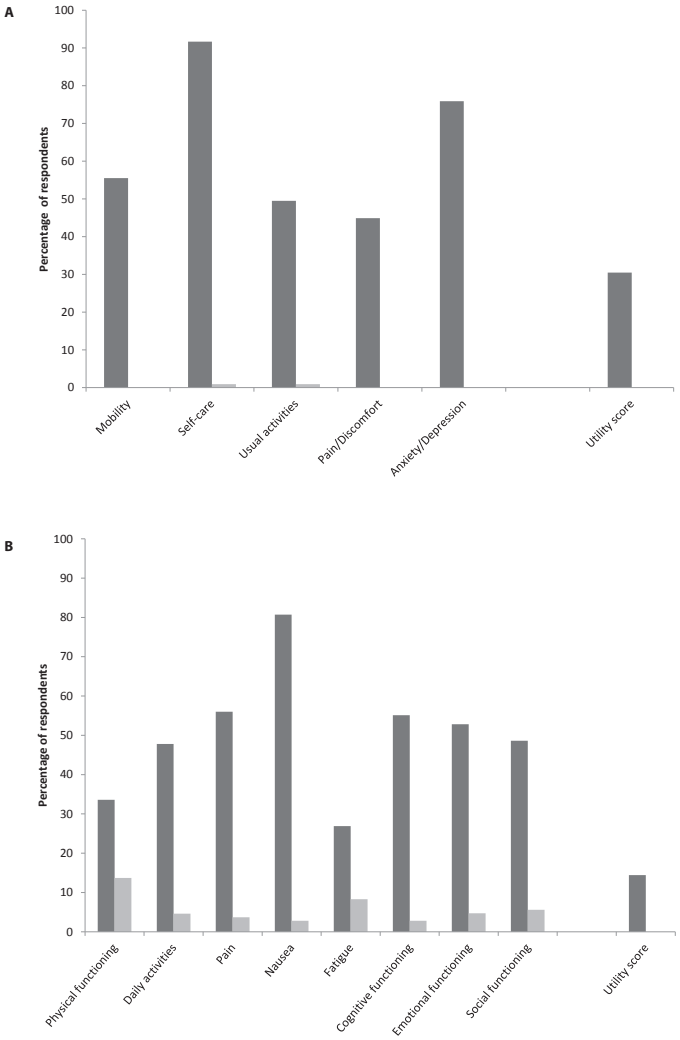


Figure 6.1 Floor and ceiling effects of the individual items and utility scores of A. the EQ-5D-5L and B. the QLQ-PBM.

The figure shows the percentage of patients with the best (dark grey) and worst scores (light gray)

score for individual items. The worst score was more frequently reported on items of the QLQ-PBM than on items of the EQ-5D-5L.

The mean utility score from the EQ-5D-5L was significantly lower than the mean utility score from the QLQ-PBM (0.83 and 0.85 respectively, p-value=0.005) (Table 6.3). In a subgroup analyses according to health status, significant differences between the utility scores of the two preference-based questionnaires were only found in the group with the poorest health status. In this group, the utility scores of the EQ-5D-5L were signifi-

Table 6.3 Utility scores for the two instruments, including the known subgroups analysis

	EQ-5D-5L		QLQ-PBM	
		P-value*		P-value*
All patients				
Mean score (SD)	0.83 (0.17)†		0.85 (0.11)	
Median score (range)	0.85 (0.21-1)		0.87 (0.51-1)	
Severity by EQ-VAS - Mean (SD)				
EQ-VAS Q1 (EQ-VAS: 0 - 65)	0.69 (0.20)†		0.76 (0.10)	
EQ-VAS Q2 (EQ-VAS: 65 - 80)	0.80 (0.14)	P<0.001	0.83 (0.06)	P<0.001
EQ-VAS Q3 (EQ-VAS: 81 - 90)	0.85 (0.14)		0.86 (0.08)	
EQ-VAS Q4 (EQ-VAS: 92 - 100)	0.96 (0.06)		0.95 (0.05)	
ES - EQ-VAS Q2 vs EQ-VAS Q1	0.65		0.91	
ES - EQ-VAS Q3 vs EQ-VAS Q2	0.33		0.38	
ES - EQ-VAS Q4 vs EQ-VAS Q3	0.97		1.25	
Severity by QL - Mean (SD)				
QL Q1 (QL: 0.0 - 66.6)	0.69 (0.18)†		0.76 (0.11)	
QL Q2 (QL: 66.7 - 83.3)	0.77 (0.15)	P<0.001	0.82 (0.06)	P<0.001
QL Q3 (QL: 83.4 - 91.6)	0.87 (0.14)		0.88 (0.09)	
QL Q4 (QL: 91.7 - 100.0)	0.96 (0.06)		0.95 (0.05)	
ES - QL Q2 vs QL Q1	0.52		0.65	
ES - QL Q3 vs QL Q2	0.65		0.77	
ES - QL Q4 vs QL Q3	0.87		0.97	
Severity by ability to work - Mean (SD) ‡				
Not able to work	0.75 (0.17)†	P=0.001	0.81 (0.11)	P=0.0003
Able to work	0.89 (0.14)		0.89 (0.09)	
ES	0.71		0.74	

QL = global quality of life scale of the QLQ-C30, ES = effect size,
*= comparison of utility scores between severity subgroups,
†= significantly different from the QLQ-PBM utility score (α=0.05)
‡= restricted to patients aged < 65 years

cantly lower than the utility scores of the QLQ-PBM. These differences can be explained by the differences in the minimum reported utility scores of the two instruments (0.21 and 0.51 for the EQ-5D-5L and QLQ-PBM, respectively).

Table 6.4 Correlation coefficients between the quality of life instruments.

		EQ-5D-5L						EQ-VAS	QL-scale
		Mobility	Self-care	Daily activities	Pain/ Discomfort	Anxiety/ Depressed	Utility score		
QLQ-PBM	Physical functioning	0.70	0.25	0.57	0.63	0.36	-0.69	-0.68	-0.68
	Role functioning	0.68	0.35	0.70	0.67	0.41	-0.77	-0.68	-0.67
	Pain	0.60	0.18	0.47	0.75	0.33	-0.70	-0.53	-0.49
	Nausea	0.21	0.20	0.20	0.09	0.26	-0.20	-0.24	-0.31
	Fatigue	0.40	0.05	0.55	0.42	0.27	-0.50	-0.59	-0.64
	Cognitive functioning	0.28	0.13	0.40	0.21	0.34	-0.36	-0.34	-0.35
	Emotional functioning	0.32	0.14	0.43	0.29	0.51	-0.45	-0.55	-0.62
	Social functioning	0.51	0.24	0.64	0.44	0.34	-0.58	-0.49	-0.56
	Utility score	-0.67	-0.28	-0.75	-0.65	-0.44	0.78	0.70	0.69
	EQ-VAS	-0.58	-0.22	-0.65	-0.53	-0.41	0.60		
	QL scale	-0.56	-0.31	-0.67	-0.55	-0.53	0.68		

The highlighted correlations indicate strong correlations (i.e. correlations larger than |0.5|). The correlations between domain scores and overall quality of life are negative, because a higher domain score indicates a worse quality of life

Both the EQ-5D-5L and QLQ-PBM utility score were strongly correlated with the EQ-VAS and QL scale (Table 6.4). Similar domains of the EQ-5D-5L and QLQ-PBM were also strongly correlated with each other. Furthermore, most domains were strongly correlated with the overall quality of life scales, with the exception of the self-care and anxiety/depression dimensions of the EQ-5D-5L and the nausea and cognitive functioning domain of the QLQ-PBM. It is expected that the weak to moderate correlations for these domains were mainly caused by the small variance in responses.

Both instruments were able to distinguish between patients with different health status, with higher utilities for poorer health status (Figure 6.2, Table 6.3). However, full health on the EQ-5D-5L was reported in all health status categories, while lower utility scores were reported by the QLQ-PBM (Figure 6.2). Although the absolute differences in QLQ-PBM utilities between health status categories were smaller than the differences in EQ-5D-5L utilities, the ES's were larger for the QLQ-PBM due to the smaller variance. Nevertheless, all calculated ES's were categorized as strong ES (larger than 0.5), except for the ES's of detecting differences between the second and third quartile of the EQ-VAS.

Reliability

Both questionnaires showed good internal consistency since Cronbach's alpha was 0.81 and 0.84 for the EQ-5D-5L and the QLQ-PBM, respectively.

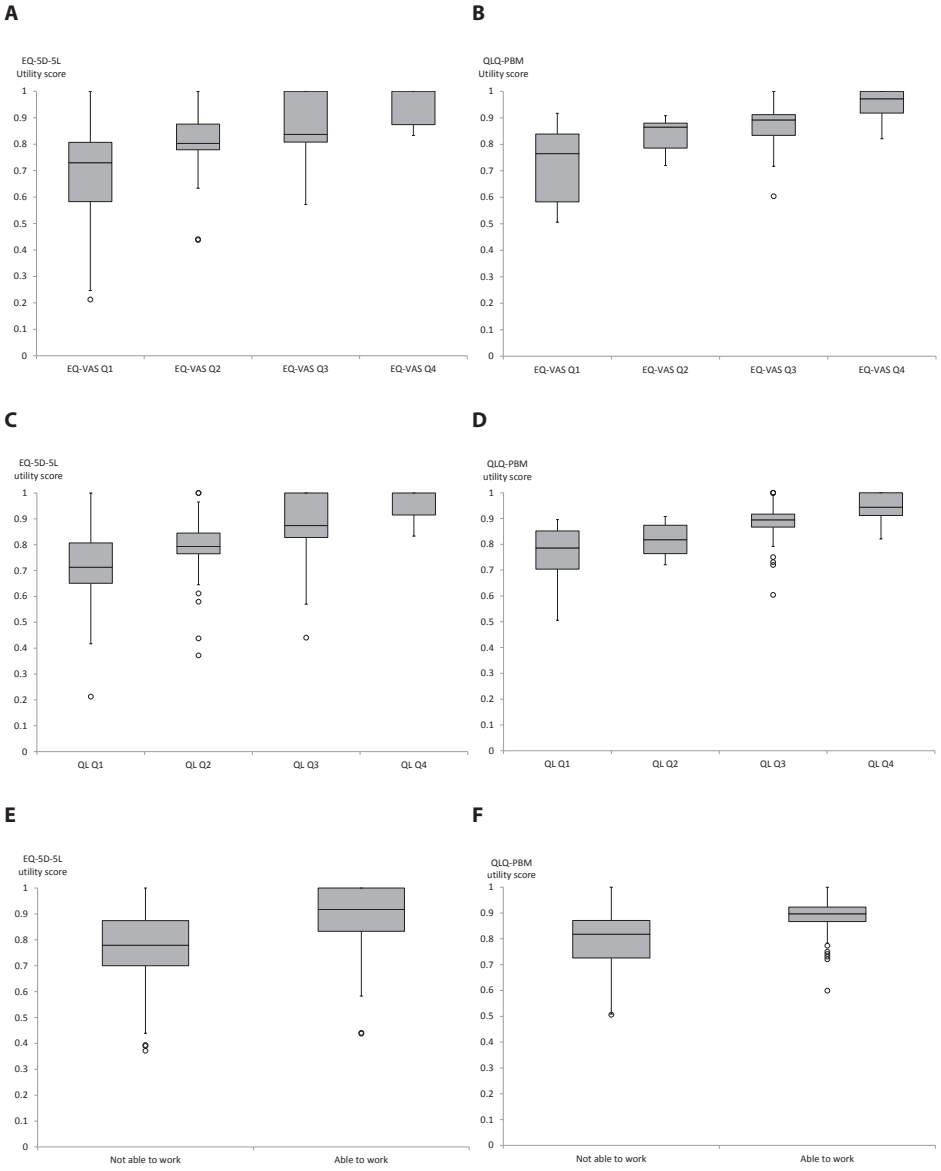


Figure 6.2 Distribution of utility scores per health status category. A. EQ-5D-5L utilities in the EQ-VAS quartiles, B. QLQ-PBM utilities in the EQ-VAS quartiles, C. EQ-5D-5L utilities in the global quality of life scale (QL) quartiles, D. QLQ-PBM utilities in the QL quartiles, E. EQ-5D-5L utilities according to ability to work, F. QLQ-PBM utilities according to ability to work.

DISCUSSION

This study showed some evidence for the feasibility, validity and reliability of both the EQ-5D-5L and the QLQ-PBM in acute leukemia patients. Both questionnaires had strong correlations with other quality of life scales and were able to discriminate between patients with different health status. Furthermore, the percentage of missing data was acceptably low for both questionnaires. The only concern about the validity is the high ceiling effect of especially the EQ-5D-5L. In total, 31% of the respondents reported full health on the EQ-5D-5L. The ceiling effect was much smaller for the QLQ-PBM (14%).

The high ceiling effect of the EQ-5D-5L might imply that the EQ-5D-5L is not able to detect all relevant differences in health as 50% of the respondents with perfect health according to the EQ-5D-5L reported problems in at least one of the QLQ-PBM dimensions. However, it can also be argued that the explicit focus on specific problems in the QLQ-PBM results in a focusing effect and thereby exaggerates the severity of these problems (105). The focusing effect might occur in this study population as the population is relatively healthy. The lowest quartiles of both the EQ-VAS and QL scale included scores up to 65 (on a 0-100 scale). Despite the relatively good health and the small differences in EQ-VAS and QL scores between the quartiles, the EQ-5D-5L was able to detect significant differences between these quartiles. Therefore, it is concluded that the high ceiling effect of the EQ-5D-5L is not problematic in this study population.

This study also showed that the differences in utility scores between health status categories were smaller for the QLQ-PBM compared to the EQ-5D-5L. Consequently, a similar improvement in health status will be valued differently depending on the type of questionnaire used to estimate utilities with smaller QALY gains and incremental cost-effectiveness ratios (ICER) for the QLQ-PBM. Similar patterns are found for other disease-specific preference-based instruments (33,156,159). In general, differences in utility scores between preference-based questionnaires can be explained by differences in the descriptive content, valuation method and scoring algorithm of the questionnaires (107).

An important difference in descriptive content between the EQ-5D-5L and the QLQ-PBM is the focus of the questionnaires. As the QLQ-PBM is specifically focused on cancer patients, more cancer relevant domains are included. For example, the QLQ-PBM includes cognitive functioning and vitality while these domains are not explicitly measured by the EQ-5D-5L. The lack of these domains in the EQ-5D is a well-known problem (159,175) and research is ongoing to evaluate whether and how the EQ-5D can be extended with these and other relevant domains (166,176-179). In contrast, the cancer-specific focus

of the QLQ-PBM might exclude the impact of comorbidities and side effects on quality of life (105,158). Other differences in descriptive content are related to differences in the formulation of items. The conceptualization of some domains differs between the EQ-5D-5L and QLQ-PBM (for example, walking around versus taking a long walk). Furthermore, differences are found in the formulation of the worst level (unable to versus many problems) and the recall period (today versus last week).

The major difference in valuation method is that the utility values of the EQ-5D-5L were not directly valued by the general public, but derived from the EQ-5D-3L (135). However, the valuation method of the EQ-5D-3L was similar to the method used in the valuation of the QLQ-PBM (87,158). It is therefore expected that the valuation method has a limited impact on the difference in utility values. Finally, differences in utility values might also result from differences in the scoring algorithm. The scoring algorithm of the EQ-5D includes an additional decrement if any of the domains is scored at the worst level, while the QLQ-PBM does not include this additional decrement (158).

The relatively good health seen amongst the patients in this study is probably a consequence of including acute leukemia survivors (172). It is assumed that survivors are relatively healthy compared to non-survivors as survivors were able to ensure both the disease and its intensive treatment. Furthermore, survivors could have adapted themselves to their new situation. Consequently, they would report relatively good health due to changes in their standards and values (response shift) (153,180). The response shift might differ between the EQ-5D-5L and the QLQ-PBM due to differences in the formulation of items. It is expected that a larger response shift would result in higher ceiling effects. Future validation studies should disentangle response shift from objective changes in health (180,181) to better interpret the differences between questionnaires.

This study has several limitations. First, the utility values for the EQ-5D-5L were based upon a cross-walk study of the EQ-5D-3L, because directly elicited preferences were not available at the time of this study. However, it is unlikely that this could have biased the results of this study since it is unlikely that directly elicited preferences would result in substantially different utilities. Another limitation of this study is that patients completed the questionnaires only once, making it impossible to evaluate the responsiveness of the EQ-5D-5L and the QLQ-PBM in acute leukemia patients. Furthermore, health status was only defined according to self-reported health status and ability to work. Although these subjective measures provide useful information about the construct validity of the questionnaires, these measures are not perfect in defining different levels of health status. It is worthwhile to assess the ability to distinguish between health status according more objective measures. Finally, the study only included acute leukemia survivors.

It might be possible that the measurement properties differ between survivors and patients receiving active treatment.

Although this study did not produce any strong evidence against the use of the EQ-5D-5L in acute leukemia, final conclusions cannot yet be drawn since this study did not cover all aspects of the validity of the EQ-5D-5L. Future studies in acute leukemia should address these gaps by studying the quality of life of patients receiving active treatment as well as assessing the responsiveness of the questionnaire over time. Furthermore, it is recommended to use a more objective measure of health status, like physician-reported ECOG performance status or relapse versus remission. As long as these additional studies confirm our findings, the EQ-5D-5L is recommended to measure quality of life utility for acute leukemia patients.

ACKNOWLEDGEMENT

The authors want to thank Bronno van der Holt from the HOVON Data Center, Erasmus MC and Bob Löwenberg from the Erasmus MC for providing clinical data of the patients who participated in this quality of life study. The authors want to thank Sarah Lonergan from the Erasmus MC for her help regarding the logistics of this study and Marjolein Schouten for her help with the data management. Furthermore, the authors want to thank Sander Arons for his valuable feedback of a draft version of this paper at a LOLA-HESG conference.

Chapter 7

The development and validation of a decision-analytic model representing the full disease course of acute myeloid leukemia

With W. Ken Redekop, Kees A.G.M. van Montfort, Bob Löwenberg and Carin A. Uyl-de Groot

Published in Pharmacoeconomics 2013(31): 605-621

ABSTRACT

The treatment of acute myeloid leukemia (AML) is moving towards personalized medicine. However, due to the low incidence of AML, it is not always feasible to evaluate the cost-effectiveness of personalized medicine using clinical trials. Decision analytic models provide an alternative data source. The aim of this study was to develop and validate a decision analytic model that represents the full disease course of AML.

We used a micro simulation with discrete event components to incorporate both patient and disease heterogeneity. Input parameters were calculated from patient-level data. Two hematologists critically evaluated the model to ensure face validity. Internal and external validity was tested by comparing complete remission (CR) rates and survival outcomes of the model with original data, other clinical trials and a population-based study.

No significant differences in patient and treatment characteristics, CR rate, 5-year overall and disease-free survival were found between the simulated and original data. External validation showed no significant differences in survival between simulated data and other clinical trials. However, differences existed between the simulated data and a population-based study.

The model developed in this study is proved to be valid for analysis of an AML population participating in a clinical trial. The generalizability of the model to a broader patient population has not been proven yet. Further research is needed to identify differences between the clinical trial population and other AML patients and to incorporate these differences in the model.

INTRODUCTION

The treatment of cancer is currently shifting towards a more personalized treatment approach. In general, treatment has always been personal because it has always been adjusted to a patient's individual characteristics and health status. However, as a result of the Human Genome Project, new methods are available to further study disease biology and variability between patients (182). Many studies have applied these new methods and found significant associations between genetic abnormalities and drug response or patient prognosis (183-187). These results provide useful information for treatment decisions. A familiar example of these new methods is found in estrogen-positive breast cancer, where multi-gene assays are able to identify patients who will benefit from chemotherapy and those who will not benefit. Consequently, the treatment choice for a patient with estrogen-positive breast cancer can depend on the results of the micro-array (188).

The personalized treatment approach also plays an increasingly important role in the treatment of acute myeloid leukemia (AML). AML is a relatively rare blood cancer caused by a proliferation of myeloid cells in the bone marrow resulting in a shortage of blood cells. The annual incidence of AML is approximately 3-4 per 100,000 (58). AML is a heterogeneous disease consisting of many (cyto)genetic subgroups with prognostic significance. Currently, patients can be classified into distinct prognostic groups based on the presence of cytogenetic and molecular abnormalities. A common way of stratifying patients according to risk considers three prognostic categories (26). Several treatment approaches are available, including various forms of high-dose chemotherapy, autologous and allogeneic hematopoietic stem cell transplantation (HSCT). Treatment choice differs between the three prognostic risk groups (21). The most hazardous treatments associated with substantial procedure related mortality, e.g. allogeneic HSCT, are restricted to patients with intermediate or high risk of relapse (189).

In recent years, new molecular subtypes of AML have been identified with new technologies. Retrospective studies have found a significant association between these new subtypes and prognosis (185,190,191). The translation of these findings into clinical practice requires an assessment of the clinical utility, which is the impact on health and economic outcomes (192). Preferably, clinical utility should be evaluated in randomized controlled trials (RCTs), because RCTs minimize bias by controlling for both known and unknown factors which might influence outcome (193). Results of RCTs evaluating the impact of the newly developed genetic tests on health outcomes are not yet available. In addition, it is questionable whether sufficiently powered RCTs can be performed in the future due to the low incidence of AML and its subtypes. Furthermore, it might be

unethical to perform an RCT and expose patients to a non-beneficial treatment if retrospective studies have shown that patients with a certain genetic deficit do not benefit from that treatment (194). The use of decision-analytic modeling provides an alternative method to assess the clinical utility of these new genetic tests in absence of RCT results. In a decision model, evidence from genomic, clinical and epidemiological data is synthesized to evaluate the risks and benefits of the new tests. In addition, decision models enable scenario analysis to evaluate the uncertainty surrounding the clinical utility (37).

Decision models used to evaluate new genetic tests should represent the full disease course from diagnosis till death in order to incorporate all relevant clinical effects of new tests. Currently available models for AML have only focused on specific parts of the disease course (195-198). Consequently, a new model representing the full disease course of AML should be developed. Therefore, the aim of this study was to develop a decision model representing the full disease course of AML from diagnosis. This paper also transparently describes the validation process of the model, which is an essential part in the development of a new decision model (199,200).

METHODS

The development and validation of the decision model was an iterative process. The complete development and validation process is shown in Figure 7.1. The first step in the development of the decision model was the design of the model structure. Once the model structure was designed, the input parameters were calculated from original patient-level data. The patient-level data consisted of a representative sample (N=427) of AML patients treated according to protocols of the Dutch-Belgian Hematology-Oncology Cooperative group (Hovon 4,29,42 and 42A, available at www.hovon.nl).(189) All of these patients were diagnosed with AML between 1987-2005. Patients received two induction chemotherapy cycles consisting of daunorubicin and an anthracycline. The treatment arms differed regarding the administration of granulocyte colony-stimulating factors and the dosing of chemotherapy. Post-remission treatment depended on the patient's risk of relapse. Favorable risk patients received high-dose chemotherapy, intermediate and unfavorable risk patients with a suitable donor received an allogeneic HSCT and the other patients received either high-dose chemotherapy or autologous HSCT.

The results of the data-analysis were discussed with clinical experts and compared with the available literature. If necessary, the data-analysis was refined according to evidence in the literature or expert opinion. The model was run once all input parameters were

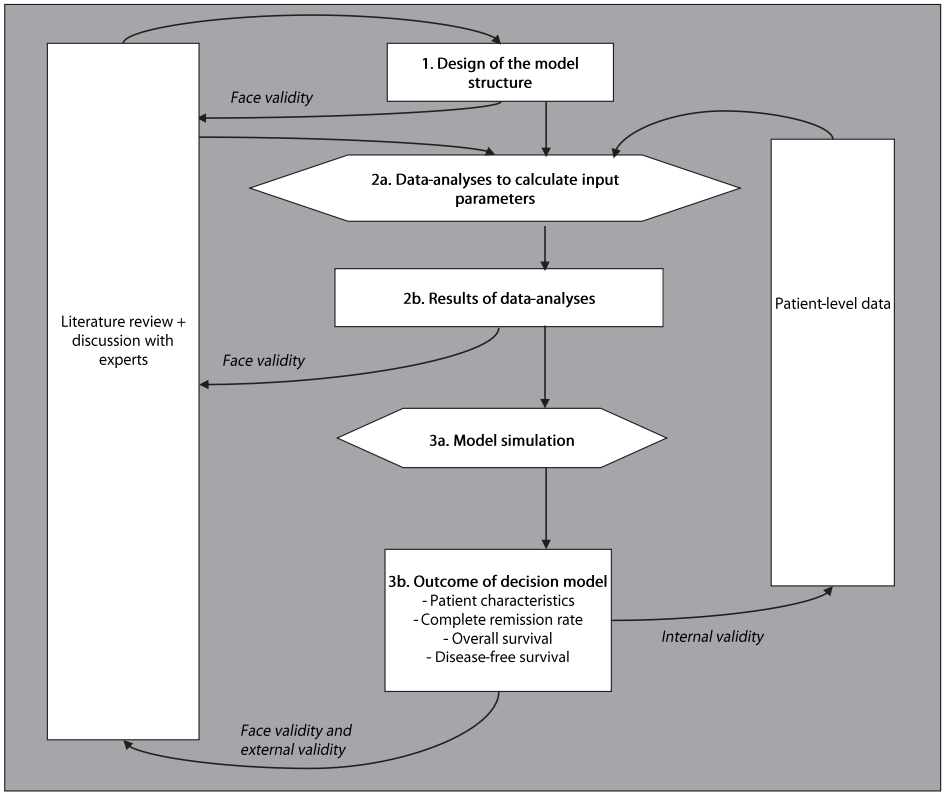


Figure 7.1 Schematic description of the development and validation process

in accordance with expert opinion and the available literature. The main results of the model were patient characteristics, complete remission (CR) rate, overall survival (OS) and disease-free survival (DFS). These outcomes were discussed with the clinical experts and compared with the original patient-level data and data sources. Adjustments were made to the model if the simulated results deviated from these data sources. All different steps are explained in more detail below.

Designing model structure

Disease characteristics

Important disease characteristics were derived from the literature and discussions with clinical experts. AML is a heterogeneous disease with a variable prognosis that can be distinguished according to a variety of patient and disease related factors at diagnosis (26,201). For instance cytogenetic abnormalities present at baseline represent a major set of prognostic factors. A risk group classification has been developed which classi-

fies patients into three risk groups according to the presence of specific cytogenetic abnormalities (64). Prognosis of AML also becomes less favorable with increasing age (13). The poorer prognosis in older patients is caused by comorbidities, different disease biology and poor tolerance of chemotherapy (201). High white blood cell (WBC) count at diagnosis and AML appearing after other hematological disorders or after previous chemotherapy and/or radiotherapy (so called secondary leukemia) are also associated with a worse prognosis (202,203).

AML treatment starts with intensive chemotherapy (consisting of cytarabine and an anthracycline) to induce a complete remission (CR) by cytoreduction of the number of leukemic blasts (64). The induction treatment consists of two chemotherapy cycles of about one month each. After each cycle, response evaluation is performed to evaluate whether a CR (less than 5% blasts in the bone marrow, disappearance of any known leukemic deposits and recovery of blood cell counts) is reached (26,204). This evaluation is scheduled at day 21 as counted from the start of the chemotherapy cycle and will be repeated if the marrow is not conclusive. All patients receive the second cycle, irrespective of the response after the first cycle. Patients without a CR after two cycles have refractory leukemia. These patients have a very poor prognosis with limited treatment options. Some patients are eligible for an allogeneic HSCT, but the majority of the patients receive investigational therapy or palliative care (198).

Patients in CR and with a good performance status receive additional post-remission treatment to further remove leukemic blasts; these options include high-dose chemotherapy, autologous and allogeneic HSCT (26,204). In general, an HSCT is not administered to patients with abnormalities which are associated with a favorable prognosis, because the benefits of an HSCT do not offset the treatment-related mortality in these patients. Patients with intermediate or unfavorable abnormalities are more likely to benefit from HSCT (189,205). Despite intensive treatment, the risk of disease relapse is still quite high, especially in patients with unfavorable abnormalities, older age and high WBC counts. In addition, patients who did not achieve a CR after 1 cycle have a higher risk of relapse (206). Prognosis of relapsed patients is poor; 5-year overall survival is 11%. Important prognostic factors for survival after relapse are duration of first remission, cytogenetic and molecular abnormalities at time of diagnosis, age at time of relapse, HSCT in first CR and achievement of a second CR (207).

Model structure

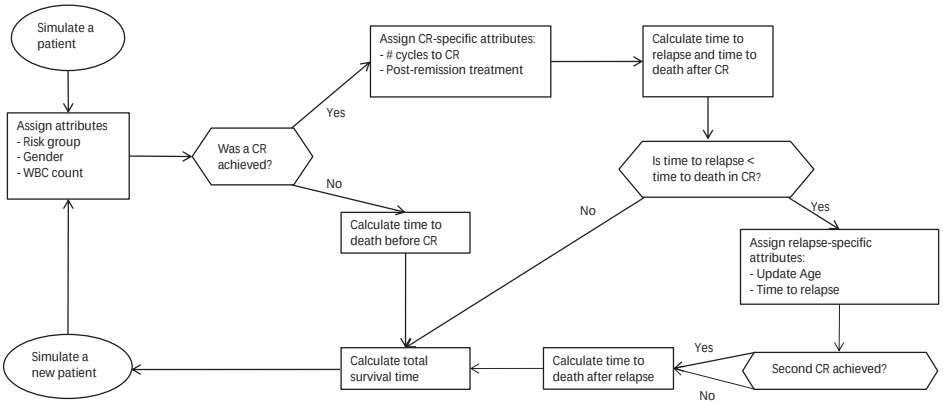
The above described disease characteristics of AML show that AML is a heterogeneous disease with several patient- and disease-related factors influencing the chance of CR and the risk of relapse and death. In addition, prognosis also depends on patient's

history regarding response to induction chemotherapy and duration of first CR. As the inclusion of all relevant prognostic factors in a cohort model would require many different health states, a micro simulation with discrete event components was considered more efficient in modeling the disease course of AML (208).

The entities in this model were patients with primary AML aged 18-60 years. The model focused on this age group, because it is expected that genetic tests provides the most value in that age group at this moment. Various treatments are available for this age group and the use of genetic tests might change treatment for specific patients due to a reclassification of the risk groups. In older patients, treatment is overall far less satisfactory and outcome following different treatments vary within a much smaller range (64). Treatment choice in older individuals is more strongly driven by the patient's performance status. Therefore, genetic tests will currently not be used to guide treatment for patients older than 60 years. The model was restricted to patients with primary AML because the number of patients with secondary AML was limited in the available database. Finally, the current model excluded patients with acute promyelocytic leukemia as these patients are treated differently than other AML patients (204).

Relevant attributes of AML patients were cytogenetic risk group, age, sex, WBC count, number of cycles needed to achieve CR, post-remission treatment and time to relapse. Three cytogenetic risk groups were identified according to the classification proposed by Döhner et al.(26): favorable, intermediate and unfavorable risk group. WBC count was dichotomized in two categories: low (WBC count < 100 x 10^9/l) and high WBC count (WBC >= 100 x 10^9/l)(26). The post-remission treatment options were no post-remission treatment, intensive chemotherapy, autologous HSCT, allogeneic HSCT from sibling donor, allogeneic HSCT from a matched unrelated donor (MUD) and umbilical cord blood (UCB) transplantation.

In general, four different events were identified for AML: CR, relapse, second CR and death. The achievement of a (second) CR was only evaluated after (re-)induction treatment. Patients were considered to have refractory disease if they had not achieved a CR after two induction cycles. Consequently, a continuous survival function for CR achievement does not provide an accurate reflection of the natural disease course. Therefore, it was first determined whether patients achieved a CR, followed by the determination of the time to CR. Two different types of deaths were identified: death due to AML and death from other causes. Both types of death could occur at different disease stages: before CR (in patients never achieving CR), after CR and after relapse.



WBC = white blood cell ; CR = complete remission

Figure 7.2 Model structure

The total structure of the decision model is described in Figure 7.2. At the start of the simulation a new patient was simulated and the attributes risk group, gender, WBC count and age were assigned to that patient. Subsequently, it was determined whether or not the patient achieved a CR. If no CR was achieved, time to death was estimated. Once a patient achieved a CR, the CR-specific attributes post-remission treatment and number of cycles needed to achieve a CR were assigned to the patient and the time to relapse and death after CR were calculated. A patient relapsed if the time to relapse was shorter than the time to death after CR. In relapse, age was updated. Subsequently, it was determined whether a second CR was achieved, followed by the calculation of time to death.

Calculation of attributes and time-to-events

Patient-level data was used to calculate input parameters for the model. First, it was evaluated whether age, gender, WBC count, risk group and post-remission treatment were correlated with each other. No significant correlations were found between gender and the other patient characteristics. The distributions for WBC count and post-remission treatment differed between the three risk groups and patients in the favorable risk group were significantly younger than patients in the other risk groups. Subsequently, frequency distributions and summary statistics of all characteristics were derived from the patient-level data. The frequency distributions for age, WBC count and post-remission treatment were derived for each risk group separately.

Logistic regression analyses were performed for achievement of (second) CR and whether a CR was achieved after the first chemotherapy cycle. All patient, disease and treatment characteristics were included as covariates in the regression models. Non-sig-

Table 7.1 Coefficients of logistic regression models and survival models (log-normal and log-logistic)

	CR	CR after 1 cycle	Death if no CR		Relapse			Death in CR		Second CR	Death in relapse	
			<=1 yr	> 1 year	<= 6 months	7-18 months	> 18 months	<= 6 months	> 6 months		<=1 year	>1 year
	(N=427)	(N=275)	(N=81)	(N=18)	(N=275)	(N=250)	(N=147)	(N=275)	(N=229)	(N=153)	(N=146)	(N=42)
Model type	Logistic	Logistic	Log-N	Log-N	Log-N	Log-N	Log-N	Log-N	Log-N	Log-N	Log-L	Log-N
Intercept	2.828	0.079	6.110	6.406	6.124	6.400	10.085	7.688	18.918	2.135	4.672	5.307
Age	-0.020	0.023	-0.032		-0.004			-0.013	-0.002	-0.044		
Favorable risk	0.278		0.053		0.562	1.380	1.544			0.398	0.734	1.113
Unfavorable risk	-1.157	-0.640	-0.021	-0.972	-0.581	-2.472	-2.694		-1.072	-0.185	-0.240	-0.920
High WBC count	-1.223				-0.325					-1.454	-0.313	-1.497
2 cycles to CR					-0.429			-0.655		-0.211		-1.713
Autologous HSCT					0.341					-0.784*	-0.159*	-1.020*
Allogeneic HSCT from sibling					0.420	1.802	1.396†		-7.004	-0.784*	-0.159*	-1.020*
Allogeneic HSCT from MUD‡					0.676	4.158	1.396†		-5.448	-0.784*	-0.159*	-1.020*
Time to relapse <= 6 months										-1.247	-0.592	
Time to relapse >18 months										0.934		
Second CR achieved											1.702	2.720
2 cycles to CR*HSCT												2.755
Shape parameter			1.527	2.118	0.580	2.238	3.243	1.254	4.846		0.487	1.732

Reference categories were: intermediate risk for cytogenetic risk group, Chemotherapy for post-remission treatment and time to relapse 7-18 months
Log-N = log-normal, Log-L = log-logistic, WBC = White blood cell, CR = Complete remission, HSCT = Hematopoietic stem cell transplantation, MUD = Matched unrelated donor
* The three transplantation types were combined in one group[207]
† The two allogeneic transplantation types were combined in one group due to the small number of patients with allogeneic HSCT from MUD
‡ Patients with an umbilical cord blood transplantation were categorized in the allogeneic HSCT from MUD group, because a separate category was not feasible due to small number of patients

nificant covariates were excluded from the model using backwards selection. However, if exclusion of a non-significant covariate resulted in higher Akaike Information Criteria (AIC) scores, this covariate was retained in the model. The final model according to the above-mentioned statistical criteria was compared with the literature about factors associated with achievement of a (second) CR and number of cycles needed to achieve CR.(13,44,209) Furthermore, the selected covariates were discussed with clinical experts. If any excluded characteristics were considered relevant according to the literature or the experts, these characteristics were added to model. Table 7.1 shows the covariates in the final models with the corresponding coefficients.

Survival analyses were performed on the patient-level data to describe the time to relapse and the time to death due to AML. For each event, a piecewise exponential regression was first performed to evaluate whether the hazard was constant over time. Different cut-off points (6 months, 1 year, 18 months) were tested by comparing the AIC of the models using these different cut-off points. The cut-off points were chosen because most deaths and relapses occur in the first two years.(64) The best fit was achieved with a cut-off point at 1 year for death if no CR and death after relapse, a cut-off point at 6 months for death in CR and a cut-off point at 6 and 18 months for relapse.

Due to the finding of the significant different hazards, different survival models were specified for each time period. Patients with a longer follow-up than the specified cut-off point were censored at the cut-off point in the analysis of the first time period. For the analysis of the second time period, survival was modeled as time since cut-off point. That analysis excluded patients with a shorter follow-up than the specified cut-off point. For all analyses, four different parametric survival models were tested (exponential, Weibull, log-normal and log-logistic distribution) and model fit was determined using the AIC score. The models with the best fit are shown in Table 7.1. Relapse and death in CR were two competing events. In the modeling of these two competing events, patients who experienced the other event were censored at time of that event. The method for selection of attributes in the logistic regression models was also used to select the attributes in the survival models. The final attributes and the corresponding coefficients are shown in Table 7.1.

The analysis for the calculation of attributes and time-to-events were performed in SAS 9.2.

Model simulation

The model calculation started with the simulation of one patient and its attributes. Input parameters used to assign the attributes are shown in Table 7.2. The assignment

Table 7.2 Base-case estimates and corresponding distributions for the attributes in the model[a]

	Base-case estimate	Distribution	Distribution estimates
Risk group			
Favorable risk group	14.5%	Dirichlet	α_1=62
Intermediate risk group	68.6%		α_2=293
Unfavorable risk group	16.9%		α_3=72
Male	51.1%	Beta	α=218, β=209
High WBC Count			
Favorable risk group	17.7%	Beta	α=51, β=11
Intermediate risk group	23.2%	Beta	α=225, β=68
Unfavorable risk group	8.3%	Beta	α=66, β=6
Age (Years)		Empirical	
Post-remission treatment			
Favorable risk group			
- no post-remission treatment	8.3%		
- chemotherapy	91.7%	Beta	α=55, β=5
Intermediate risk group			
- no post-remission treatment	13.6%	Dirichlet	α_1=33
- chemotherapy	48.6%		α_2=118
- autologous HSCT	14.8%		α_3=36
- allogeneic HSCT from sibling	23.0%		α_4=56
Unfavorable risk group			
- no post-remission treatment	18.8%	Dirichlet	α_1=9
- chemotherapy	18.8%		α_2=9
- autologous HSCT	8.3%		α_3=4
- allogeneic HSCT from sibling	35.4%		α_4=17
- allogeneic HSCT from MUD	16.7%		α_5=4
- UCB transplantation	2.1%		α_6=17

of the categorical attributes was based on the comparison of the predefined frequency distributions with a random number between 0 and 1 which was drawn from a uniform distribution. A separate random number was drawn for each categorical attribute. Age was derived from an empirical distribution. This distribution differed between favorable risk and other patients.

After the assignment of the attributes at start of simulation, the personalized chance of CR achievement was calculated with formula (1):

$$\frac{e^{(\beta_0+\beta_1X_1+\cdots+\beta_nX_n)}}{1+e^{(\beta_0+\beta_1X_1+\cdots+\beta_nX_n)}} \quad (1)$$

The relevant beta coefficients are shown in Table 7.1. If a patient achieved a CR, the next step was to determine whether the patient achieved a CR after one cycle or not. The probability of achieving a CR after one cycle was also calculated with formula (1) using the beta coefficients shown in Table 7.1. The time to CR was derived from an empirical distribution which was observed in the original patient-level data. Two different distributions were developed: one for CR after one cycle and one for CR after two cycles.

The times to death due to AML and relapse were derived the survival functions shown in Table 7.1[1]. Time to death from other causes was derived from life expectancy estimates of Statistics Netherlands according to the patient's age and gender (72). For patients in CR, the subsequent event was determined by the shortest time-to-event; if the time to relapse was shorter than both times to death, the patient relapsed. If a patient relapsed, it was determined whether the patient achieved a second CR with formula (1) and the coefficients shown in Table 7.1. Subsequently, times to death were estimated. Death was considered as the only possible event for patients without CR and patients in relapse. Cause of death was determined by which of the two times to death (time to death due to AML or time to death from other causes) was shorter.

Finally, the OS and DFS time were calculated for each simulated patient. OS was measured as time from diagnosis till death and DFS was measured as time from CR till relapse or death in CR.

In order to test the validity and reliability of the model, it was decided to simulate a sample size equal to the sample size in the original data (N=427). Subsequently, the simulation process was repeated 200 times to ensure stable estimates of CR rates, 5-year OS and DFS. The same input parameters were used in all 200 iterations. This analysis is called the base-case analysis.

Probabilistic sensitivity analysis (PSA)

A second order Monte Carlo simulation was used to perform a probabilistic sensitivity analysis to assess the impact of parameter uncertainty. In total, 250 different runs were modeled using different input parameters in each run. These parameters were randomly drawn from predefined distributions. Beta distributions were used for bivariate variables,

1. Lognormal distribution: log(survival time) is normally distributed with mean= $\beta_0+\beta_1X_1+\ldots+\beta_nX_n$, and standard deviation= shape parameter
Loglogistic distribution: survival time = scale * $(p/(1-p))^{(1/\text{shape})}$, where scale = $\exp(\beta_0+\beta_1X_1+\ldots+\beta_nX_n)$

Dirichlet distributions for multivariate variables and multivariate normal distributions for coefficients of logistic regression and survival models (Table 7.2). Each run consisted of 200 iterations in which 427 patients were simulated.

The decision model was programmed in Excel 2010 with macros programmed in Visual Basic for Applications.

Model validation

Three different approaches of validation were used to test the validity of the decision model (199,200,210). During the development process, the decision model was continuously discussed with two hematologists(BL and PS) to ensure face validity. BL is a specialist in the field of acute myeloid leukemia and PS is the head of the hematology department of the academic hospital in Rotterdam. Regular meetings were scheduled with these clinical experts. These meetings were scheduled in the beginning of the development process, once the model has been structured, once input parameters were estimated and once results were available regarding simulated survival outcome. During the meetings all steps of the development process were presented in detail. The experts critically assessed the composition and results of the model. In addition, the experts had to answer some specific questions about the model. For example, the experts were asked whether all relevant prognostic factors were included in the model, how cytogenetic abnormalities should be classified and whether patients with a second CR had a prognosis that was similar to that of patients with a first CR. The specific impact of the prognostic factors on the different input parameters were also discussed with the experts. For each input parameter, the selected prognostic factors were presented to the experts. The experts assessed whether these factors were indeed relevant prognostic factors in their opinion and whether some other important factors were excluded. In most instances, the clinical experts agreed with each other about relevant health states and prognostic factors for AML. Some contentious points were related to the inclusion of prognostic factors for the specific outcome measures. In case of contentious point, BL's opinion was considered paramount as he was most experienced in AML. If necessary, the model was adjusted according to the commentary of the clinical experts. The final outcomes of the decision model in terms of CR rate and survival were also discussed with the clinical experts.

Second, internal validation of the model was performed by comparing the patient characteristics and clinical outcomes of the decision model with the original data. In the base-case analysis, patient characteristics and clinical outcomes were described with a mean value over the 200 iterations and a 95% confidence interval using the standard deviation of the 200 iterations as standard error of the mean. The estimates in the original

data were interpreted as the true value, and therefore it was evaluated whether these true values fell within the 95% confidence intervals of the model outcomes.

The external validation of the model was the last validation approach of this study. First, we considered whether the model could also be applied to other countries by comparing the estimated 5-year OS from the simulated dataset with 5-year OS results from clinical trials performed in other countries. Clinical trials were selected if the patient population consisted of patients aged 18-60 years and if the results were published between 2007 and 2012. Secondly, we considered whether the model results could be generalized to the total AML population by comparing the OS estimates from the base-case analysis with population-based OS estimates. These population-based estimates were derived from the national cancer registry and included all patients diagnosed with AML in the Netherlands between 2001 and 2010.(211) More specific information regarding risk group classification and received treatment was not available for this dataset.

RESULTS

Internal validation

Table 7.3 shows the patient and treatment characteristics in both the original and simulated data, including the 95% confidence interval for the simulated data. The simulated characteristics were almost all identical to the values in the original data, with the exception of post-remission treatment and time to relapse. A slightly higher, but not significant, percentage of patients received high-dose chemotherapy in the simulated data. However, the treatment distribution within the three risk groups did not differ significantly between the simulated and original data (data not shown). The difference in the total treatment distribution might be related to differences in the risk group distribution of patients in CR due to the exclusion of patients who were not treated according to protocol in the original data. Although no significant differences were found regarding the timing of relapse, there was a trend towards fewer relapses in the first 6 months and more relapses after 18 months in the simulated data.

CR rates for the total population were similar in the original and simulated data (Table 7.4). The survival curves of the two datasets shows that the simulated OS and DFS were slightly higher compared to the original data (Figure 7.3). However, these differences were not significant since the estimates of the original dataset fell within the 95% CI of the simulated results. Subgroup analysis regarding the three cytogenetic risk groups also showed no significant differences in CR rates, OS and DFS between the simulated and original dataset (Table 7.4 and Figure 7.4). The inclusion of uncertainty surrounding

Table 7.3 Patient and treatment characteristics

	Original data		Simulated data	
	Mean	95% CI	Mean	95% CI
All patients				
Mean age (years)	43.1	(42.0 – 44.2)	43.1	(42.0- 44.1)
Gender (%)				
Male	51.1	(46.2 – 55.9)	51.1	(47.1 - 55.1)
Female	49.0	(44.1 – 53.8)	48.9	(44.9 – 52.9)
White blood cell count (%)				
<= 100 x 10^^9/l	80.1	(76.0 – 83.8)	80.1	(76.4 - 83.8)
> 100 x 10^9/l	19.9	(16.2 – 24.0)	19.9	(16.2 – 23.6)
Cytogenetic risk group (%)				
Favorable cytogenetics	14.5	(11.3 – 18.2)	14.4	(10.9 – 17.9)
Intermediate cytogenetics	68.6	(64.0 – 73.0)	68.8	(64.3 - 73.3)
Unfavorable cytogenetics	16.9	(13.4 – 20.8)	16.8	(13.1 - 20.6)
Patients with a complete remission				
Number of cycles before CR (%)				
1 cycle	72.3	(67.2 – 76.9)	72.1	(67.0 – 77.1)
2 cycles	27.8	(23.1 – 32.8)	27.9	(22.9 – 33.0)
Post-remission treatment (%)[a]				
High-dose chemotherapy	47.7	(42.0 – 53.1)	51.2	(45.8 - 56.5)
Autologous HSCT	12.4	(9.0 – 16.5)	11.4	(8.2 – 14.7)
Allogeneic HSCT from sibling donor	22.6	(18.2 – 27.6)	21.2	(17.0 – 25.4)
Allogeneic HSCT from MUD	2.5	(1.1 – 4.8)	2.4	(0.9 – 3.8)
Cord blood transplantation	0.3	(0.0 – 1.7)	0.3	(0.0 – 0.9)
Patients with relapse				
Time of relapse (%)*				
<= 6 months after CR	30.6	(23.9-38.1)	25.7	(18.9 – 32.4)
7-18 months after CR	53.2	(45.4-60.8)	53.1	(46.2 – 60.0)
> 18 months after CR	16.2	(11.0-22.5)	21.2	(14.9 – 27.6)

* patients who were not treated according to protocol were excluded in the original patient-level data (N=34)
CR = complete remission

the input parameters resulted in considerable ranges of CR rates and 5-year OS and DFS estimates, especially in the favorable and unfavorable risk group (Table 7.4).

External validation

In total, 6 articles were selected which reported 5-year OS results from clinical trials performed in several European countries, the United States, Korea and Japan (212-217). A summary of these studies is shown in Table 7.5. The simulated study is based on patients

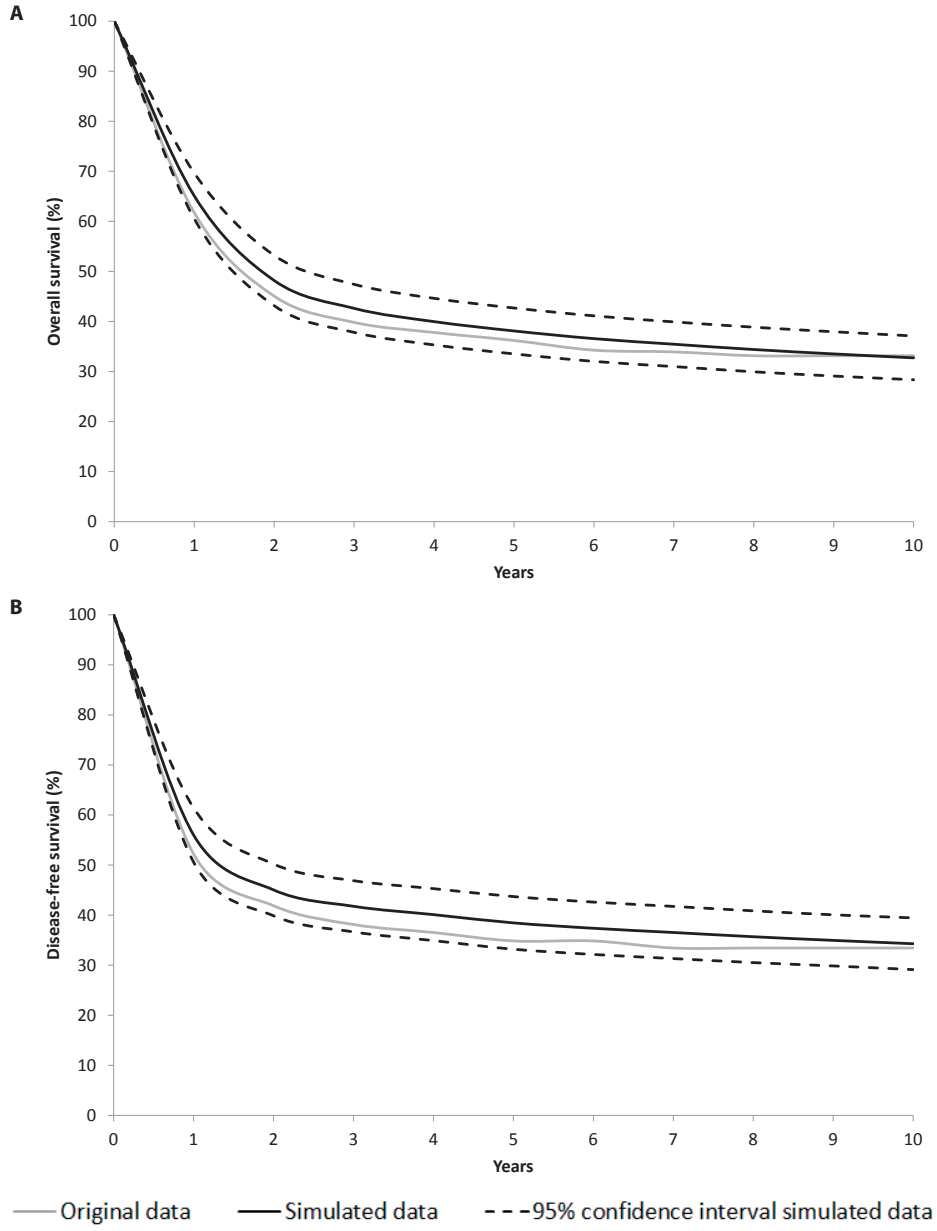


Figure 7.3 Comparison of overall and disease-free survival between original and simulated data

diagnosed from 1987-2005, while most selected trials included patients diagnosed since 2001. Only Mandelli et al. (214) and Wheatley et al. (216) reported the results of patients diagnosed in the 1990s. The median age in the studies ranged from 43-49 years. The two studies in the UK (212,216) were not restricted to patients aged 15-60 years. All but one

Table 7.4 Complete remission rates, 5-year overall and disease-free survival in original and simulated data

	Original data		Base-case analysis		Probabilistic analysis	
	Mean	95% CI	Mean	95% CI	Mean	Range
Events						
First complete remission (%)						
- All patients	81	(77 – 85)	81	(77 - 85)	81	73-85
- Favorable risk group	89	(78 – 95)	89	(80 - 97)	88	69-96
- Intermediate risk group	83	(78 – 87)	83	(79 - 87)	83	73-88
- Unfavorable risk group	67	(55 – 77)	67	(56 - 78)	65	50-80
Second complete remission (%)						
- All patients	39	(32 – 47)	41	(34 - 48)	41	28-50
- Favorable risk group	56	(28 – 85)	63	(39 - 88)	62	29-88
- Intermediate risk group	40	(32 – 50)	40	(32 - 49)	40	26-51
- Unfavorable risk group	28	(13 – 46)	32	(17 - 46)	32	13-58
5-year overall survival (%)						
- All patients	36	(31 – 41)	38	(34 - 43)	38	27-46
- Favorable risk group	65	(51 – 79)	66	(54 - 78)	64	43-77
- Intermediate risk group	37	(31 – 43)	38	(33 - 44)	38	28-47
- Unfavorable risk group	16	(7 – 24)	14	(6 - 22)	14	6-30
5-year disease-free survival (%)						
- All patients	35	(30 – 40)	38	(33 - 44)	37	25-44
- Favorable risk group	61	(45 – 77)	62	(50 - 74)	60	42-75
- Intermediate risk group	35	(29 – 41)	38	(32 - 44)	36	24-47
- Unfavorable risk group	15	(4 – 25)	15	(6 - 25)	13	3-27

95% CI = 95% confidence interval

study (214) used a risk-adapted treatment protocol with more intensive treatment for patients in the intermediate and unfavorable risk group.

According to patient and treatment characteristics, the simulated study was most comparable with the trials from Burnett et al. (212) and Lee et al. (213) The 5-year OS in these studies was slightly higher than the 5-year OS in the simulated study (41-43% and 38% respectively). A possible explanation for the difference in 5-year OS is the time of patient accrual. A trend for a better OS in more recently diagnosed patients was also found in the original patient-level data on which the simulation is based. Despite comparable treatment protocol, it might be possible that the better OS in more recently diagnosed patients is caused by improvement in supportive care (212). This is also supported by the fact that the 5-year OS in the two other studies with patients diagnosed since 2001 was similar or higher than the 5-year OS in the simulated study (215,217). The similar

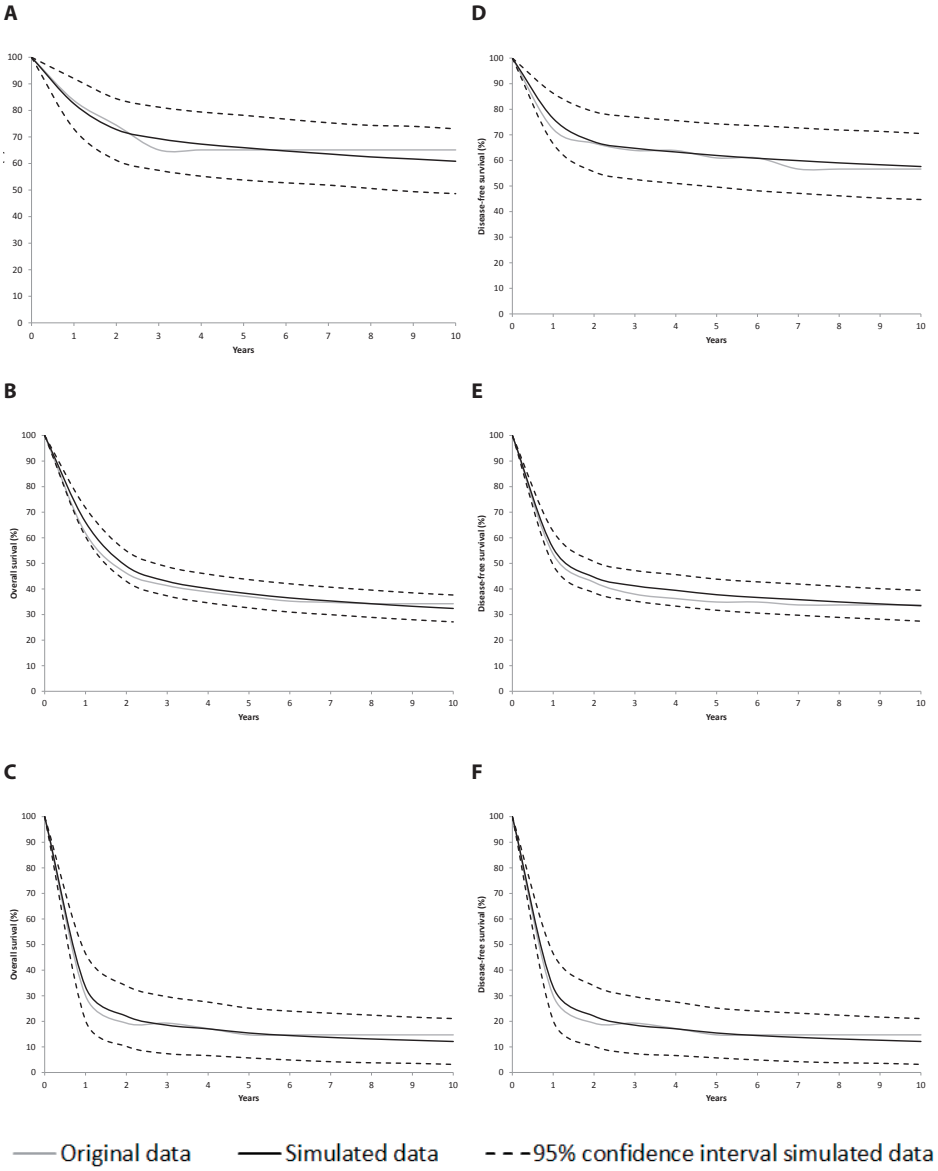


Figure 7.4 Comparison of overall survival (OS) and disease-free survival (DFS) between original and simulated data per cytogenetic risk group (A and D = favorable risk group, B and E = intermediate risk group, C and F = unfavorable risk group).

survival might be caused by the fact that an allogeneic HSCT was not a standard treatment option in that study.

The 5-year OS in the simulated study varied largely from the 5-year OS in three studies (214-216). The two studies with worse survival included patients diagnosed in the 1990s. Furthermore, treatment might also be responsible for difference in OS. In one study, a second induction cycle was only administered to patients with partial remission, while in the simulated study the second cycle was administered to all patients. This might have resulted in a lower CR rate in that study (70% versus 81%) and consequently a worse 5-year OS. In the other study, the 5-year OS was only lower in the arm in which patients received G-CSF. A better OS was observed in the Japanese study (215). However, that trial included a patient population with a relatively better prognosis as fewer patients were classified in the unfavorable risk group.

Figure 7.5 shows the survival in both the simulated and population-based data in three separate age groups. The model only simulates comparable survival in patients aged 45-54 years. The simulated survival was worse in the younger age group and better in the older age group. An interpretation of these differences is difficult due to the limited information of the population-based data.

DISCUSSION

The aim of this study was to develop a decision model for AML which can be used to evaluate the clinical utility and cost-effectiveness of genetic tests for AML patients in order to facilitate the decision-making about the widespread implementation of these tests. The decision model developed in this study included patient and disease heterogeneity in a micro simulation with discrete event components. The modeled OS was comparable with the estimates found in original patient-level data and other clinical trials. However, the generalizability to a broader patient population has not been proven. It can therefore be concluded that the developed decision model validly represents a clinical trial AML population, but can currently not be used in a broader population. Although the model was specifically developed to evaluate genetic tests, its use is not only restricted to the evaluation of these tests. Since the model represent the full disease course of AML from diagnosis, it is also feasible to use the model in cost-effectiveness analyses of new chemotherapies or other treatments.

This study also provides a transparent example of a validation process. Although guidelines about decision-analytic modeling state that the validity of the model should be

Table 7.5 Summary of the studies used in the external validation phase

Study	Country	N	Patient accrual (years)	Median age (range) in years	Risk group classification	Induction treatment	Post-remission treatment	5-year overall survival
Burnett et al. (2011)	United Kingdom	1,113	2002-2006	49 (0-71)	GO: Fav 15.8%, Int 68.8%, Unfav 15.4% No GO: Fav 14.1%, Int 72.0%, Unfav 13.9%	2 cycles of either ADE, DA or FLAG-IDA. randomization between GO or no GO	Fav: 2 or 3 cycles HD chemo Int: Sibling donor: allogeneic HSCT, others: 2 or 3 cycles HD chemo Unfav: Sibling or unrelated donor: allogeneic HSCT, others 2 or 3 cycles HD chemo.	GO: 43% No GO: 41%
Kolitz et al. (2010)	United States	302	2001-2003	46 (17-59)	Favorable 15% Intermediate 68% Unfavorable 17%	ADE or ADEP, 2nd cycle for patients without CR	Fav: 3 cycles HD chemo Other: autologous HSCT or 3 cycles HD chemo Small number of patients: allogeneic HSCT (off protocol)	37% (both arms)
Lee et al. (2011)	Korea	383	2001-2008	43 (15-60)	Favorable 21.1% Intermediate 64% Unfavorable 14.9%	DA (either standard or high dose daunorubicin). 2nd cycle for patients without CR	Fav: 4 cycles HD chemo + maintenance treatment Int + Unfav: if appropriate donor: allogeneic HSCT. Others: HD chemo + maintenance treatment	All patients: 40.8% Fav: 57.2% Int: 41.8% Unfav: 14%
Mandelli et al. (2009)	Several European countries	2,157	1993-1999	44 (15-60)	Favorable 20.6% Intermediate 66.8% Unfavorable 12.5%	Cytarabine + etoposide + one of the following: DNR, MXR, IDA. 2nd induction in case of partial remission	Younger patients with sibling donor: allogeneic HSCT. Others: autologous HSCT	DNR arm: 31.4% MXR arm: 33.7% IDA arm: 34.3%
Ohtake et al. (2005)	Japan	1,064	2001-2005	47 (15-64)	Favorable 24% Intermediate 67% Unfavorable 9%	Cytarabine + either IDA or DA. 2nd cycle if no CR achieved	Randomization: 4 cycles conventional chemo-therapy or 3 cycles high-dose chemotherapy. After chemotherapy: allogeneic HSCT for patients aged < 50 years in int or unfav risk group	48% (both arms)
Wheatley et al. (2009)	United Kingdom	803	1994-1997	49 (15-77)	Favorable 19% Intermediate 67% Unfavorable 14%	2 induction cycles of ADE, MAE, DAT or MAC. Randomization between G-CSF or not	2 or 3 further courses of intensive treatment with either chemotherapy or HSCT as last course	No G-CSF: 36% G-CSF: 29%

GO = gemtuzumab ozagamicin, Fav = favorable cytogenetic risk group, Int = intermediate cytogenetic risk group, Unfav =Unfavorable cytogenetic risk group, ADE = cytarabine + daunorubicin + etoposide, DA = daunorubicin + cytabine, FLAG-IDA = fludarabine + cytarabine+ granulocyte colony-stimulating factor + idarubicine, HD = high-dose, HSCT = hematopoietic stem cell transplantation, ADEP = cytarabine + daunorubicin + etoposide + PSC-883, CR = complete remission, DA = daunorubicin + cytarabine, DNR = daunorubicin, MXR = mitoxantrone, IDA = idarubicin, MAE = mitoxantrone + cytarabine + etoposide, DAT = daunorubicin + cytarabine + thioguanine, MAC = mitoxantrone + cytarabine, G-CSF = growth colony-stimulating factor

tested, no explicit guidance is provided regarding the suitable methods for validation (199,218). In our validation process, we interpreted the CR and survival rates in the original data as true values and evaluated whether the simulated rates differed significantly from these true values. None of the patient characteristics and outcome measures differed significantly from the true values. However, the simulated outcomes measures were not always identical to the original values. It can be argued that these differences might have important implications for cost-effectiveness analyses. For example, the percentage of patients with a relapse in the first 6 months after CR was lower in the simulated data, while a higher percentage of patients had a relapse after 18 months. This means that the average time to relapse is longer in the simulated data which might results in differences in resource use, quality of life and survival. The slightly better OS and DFS in the simulated data might be a consequence of the problems in estimating the time of relapse. Another explanation for the better OS and DFS is that the simulated dataset consists of relatively more favorable risk patients.

The large range in outcome measures which was found in the PSA is not surprising as the input parameters were derived from a relatively small patient population. Especially the favorable and unfavorable risk group contained a small number of patients. However, the average results of the PSA were comparable with the original data. If the model will be used in future cost-effectiveness analyses, a PSA should definitely be performed as it has been shown that changes in the input parameters have a significant impact on the

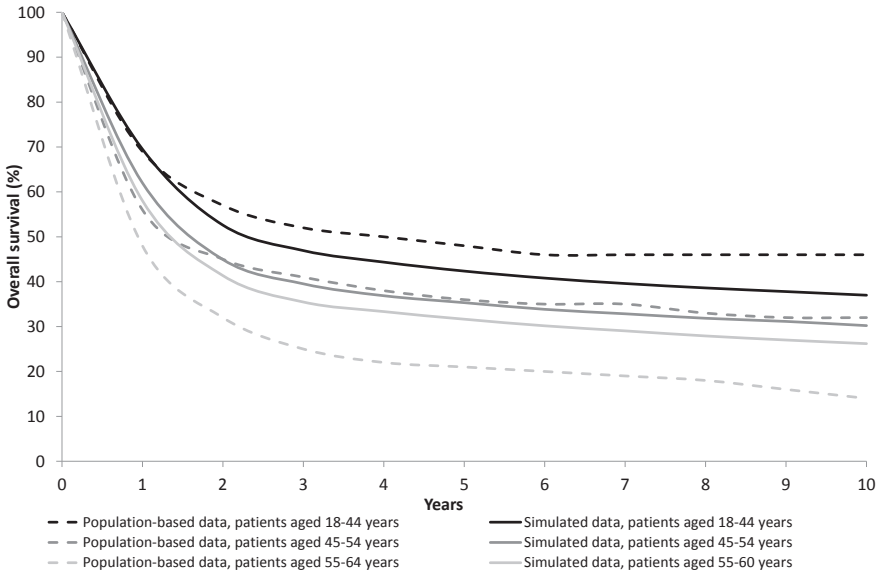


Figure 7.5 Comparison of overall survival in simulated dataset with survival in population-based data

outcome measure. Furthermore, it might be useful to perform a value-of-information-analysis to assess whether it is valuable to collect more information regarding the input parameters.

The added value of this model compared to previously existing models in the field of AML is that the scope of previous models was either restricted to post-remission treatments or to the treatment of specific complications of AML (i.e. fungal infections) (195-198). One could argue that instead of developing a completely new model, we could have modified the existing models to cover the full disease course of AML. However, some important aspects of AML were not included in the previously developed models. For example, the two studies that focused on post-remission treatments only evaluated allogeneic transplantation compared to no allogeneic transplantation. No distinction was made in type of allogeneic transplantation or between chemotherapy and autologous transplantation. In addition, not all relevant prognostic factors were included in the decision models. While Kurosawa et al. (196) used different transition probabilities for the three cytogenetic risk groups, they did not take into account the prognostic value of age, WBC count or number of cycles needed to achieve CR.

The main reason for choosing a micro simulation instead of cohort simulation was the heterogeneous nature of the disease. Several patient and disease characteristics have an impact on prognosis. This heterogeneity will only grow in the future as many studies are now trying to identify new (genetic) markers with prognostic value. This trend is not only apparent in the field of AML, but is also seen in other disease areas (188,219,220). Although a micro simulation was considered necessary because of the need to incorporate the heterogeneity of the disease, the use of discrete event components was in retrospect not essential; a micro simulation Markov model could also have been developed. (208) However, we do not believe that the model choice will influence model outcomes, because no differences in model outcomes have been found in direct comparisons between Markov cohort models and discrete event simulations (DES) (221,222).

Since models represent a simplification of the real world, it is never feasible to include all differences found in the real world. An important simplification in our model is that graft-versus-host-disease (GVHD), a complication of an allogeneic transplantation, was not explicitly modeled. However, the consequences of GVHD were implicitly included in the estimation of life expectancy after allogeneic transplantations. If a new AML treatment particularly influences the incidence or consequences of GVHD, it could be important to model GVHD explicitly.

Another simplification was made in the analysis of the two competing events 'death in CR' and 'relapse'. These two events were analyzed separately. Patients who experienced the event not of interest in that specific analysis were censored at time of the experienced event. However, an important assumption here is that censored patients do not differ from non-censored patients. One can doubt whether this assumption holds in this specific situation as there might be differences between patients who die in CR and patients who relapsed. However, since the model results corresponded well with the original data, it was not considered necessary to use more sophisticated methods in the decision model.

The impact of the structural uncertainty of the model has not explicitly been evaluated in this study. It was assumed that the chosen model structure reflects reality as clinical experts have confirmed that structure. A limitation of the reliance on these two experts is that they were both located in the same hospital. Their opinions might therefore not necessarily represent (inter)national opinions. However, this lack of generalizability is likely to be minimal as the experts collaborate with different centers all over the world. (223,224) Furthermore, no large differences were found between the modeled 5-year OS and the 5-year OS reported in comparable clinical trials conducted in other parts of the world (212-214,216,217). Nevertheless, it is recommended to discuss the model structure with other experts in the future and, if necessary, assess the impact of changes in structure on the survival outcomes.

Another limitation of this study is that all input parameters are derived from one relatively small study. This is further complicated by the fact that the hazard of death and relapse was not constant over time. As a result, some distributions are based on a small group of patients who experienced the specific event. Due to the small numbers, it might be possible that relevant effects have been missed. Furthermore, a trend was found for an improved survival in more recently diagnosed patients Due to the small number of patients in the total study sample, it was decided to also include patients who have been diagnosed more than 15 years ago. However, the model might be improved if more recent data of a substantial study sample were to be included.

The generalizability of the model to a broader patient population than patients participating in clinical trials has not been proven in this study. Possible explanations for the deviations between the simulated study and the population-based data might be related to the patient characteristics or the time of diagnosis. Further research is required to identify differences between study populations and to adjust the model to ensure generalizability to a broader population.

CONCLUSIONS

This study showed a new valid decision model for AML patients in a clinical trial setting. The existence of a validated model facilitates assessments of the clinical utility and cost-effectiveness of new genetic tests for AML. Since it is not often clear how genetic tests will alter disease management, the model can also be used to evaluate the impact of different scenarios regarding treatment change in newly identified subgroups of AML on survival, quality of life and costs. Furthermore, the model can also be used for other cost-effectiveness analysis in the field of AML.

ACKNOWLEDGEMENTS

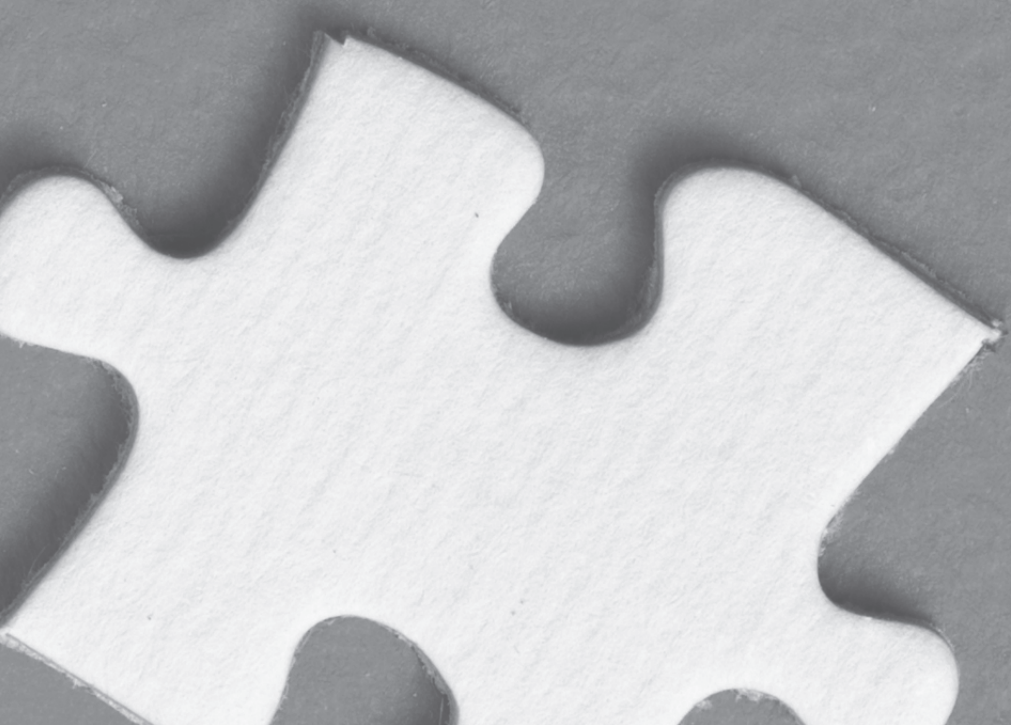
The authors want to thank Professor Pieter Sonneveld for his expert advice in the validation of the decision model and Wim van Putten from the Hovon Data Center for providing access to patient level data and his statistical advice during the development of the model.

Chapter 8

Methodological recommendations for cost-effectiveness analyses of personalized medicine strategies

With W. Ken Redekop, Bob Löwenberg and Carin A. Uyl-de Groot

Submitted



ABSTRACT

An adequate assessment of the health and economic consequences of personalized medicine strategy is crucial for the implementation of personalized medicine strategies in clinical practice. Different methods are currently used in cost-effectiveness analyses of personalized medicine strategies due to the absence of specific methodological recommendations. The aim of this study was to propose specific methodological recommendations for cost-effectiveness analyses of personalized medicine strategies based upon findings from a case-study in acute myeloid leukemia (AML).

A previously validated decision-analytical model was used to assess the cost-effectiveness of two personalized medicine strategies in AML patients aged 18-60 years. These strategies include the reclassification of patients to either the favorable or unfavorable risk group and the subsequent treatment change. The cost-effectiveness analyses were first restricted to an assessment of the consequences of a treatment change in the newly identified subgroups. Subsequently, the consequences of the test were added to the analyses. The impact of the exclusion of the test was assessed by comparing the results of the two approaches.

We found that the exclusion of the testing costs in cost-effectiveness analyses overestimates the cost-effectiveness of personalized medicine strategies. The severity of overestimation depends upon the costs of the test, the prevalence of the subgroup with a treatment change and health effects of the treatment change. According to this findings, we developed a formula for use in future cost-effectiveness analyses of personalized medicine strategies.

This study showed that cost-effectiveness analyses of personalized strategies should always include the costs and performance of the test to avoid an overestimation of the cost-effectiveness. Nonetheless, insight in the separate effect of the treatment and test on health and economic outcomes is required for adequate reimbursement decisions. This can be achieved by using the postulated formula in future cost-effectiveness analysis.

INTRODUCTION

A general trend in the field of medicine is the personalization of medical treatment. As it is known that treatment response and side effects differ between patients,(8) many studies focus on the identification of the underlying causes of these differences (225,226) to improve health outcomes by treating patients according to their expected response and risk of side effects (227).

Despite the promising results of personalized medicine strategies, only a small number of personalized medicine strategies have been implemented in daily practice. A major barrier of implementation is the limited evidence of the impact of personalized strategies on health outcomes and costs (228-230). Direct evidence from randomized controlled trials (RCTs) is often missing since that information is not specifically required for market access, except for companion diagnostics (228).

Furthermore, RCTs are not always feasible due to ethical reasons and time and costs constraints (194). Therefore, the available evidence regarding impact is often derived from retrospective studies using modelling methods to combine different sources (37,194,231,232). A disadvantage of the latter approach is that it furnishes indirect evidence and differences in health outcomes and costs cannot be attributed to differences in treatment with certainty (233).

Another difficulty in the successful implementation is the lack of standardized methods in the evaluation and reporting of cost-effectiveness analyses of personalized medicine strategies. Many published cost-effectiveness analyses excluded the costs of testing and only evaluated the consequences of a new treatment in a narrowly defined subgroup of patients (234,235). Other studies performed cost-effectiveness analyses of the combination of the test and treatment strategy (236-238). The results of the latter studies are often not comparable with the more narrow cost-effectiveness analyses as the separate effect of the test and treatment on health outcomes and costs is not reported.

Specific recommendations regarding the appropriate methodology for cost-effectiveness analyses of personalized medicine strategy are currently not available. Detailed cost-effectiveness guidelines were developed for the assessment of new drugs and lack information about the evaluation of the test (29,239,240). Nonetheless, several authors indicated that the method for cost-effectiveness analyses of drugs also apply to analyses of personalized medicine strategies. The analyses should only be expanded with an assessment of the costs and effects of the test strategy (241,242). However, none of these

authors explicitly describe the required adjustments in the methodology to allow for the assessment of both the test and treatment strategy.

It is expected that stricter recommendations regarding the methodology of cost-effectiveness analyses of personalized medicine strategies improves the comparability between studies. A better comparability may improve the judgments about the added value of new personalized medicine strategy. Since new personalized medicine strategies may incorporate adjustments in both the test and treatment, it is essential that the methodology allows for the separate assessment of the test and treatment strategy. The aim of this study was to examine whether specific methodological recommendations for cost-effectiveness analyses of personalized medicine strategies could be proposed based upon findings from a case-study in acute myeloid leukemia.

METHODS

Description case study

Acute myeloid leukemia (AML) represents a highly heterogeneous disease in terms of clinical and molecular features and disease outcome. It offers an interesting case study for personalized medicine due to its heterogeneous nature and the involvement of various different cytogenetic and molecular subgroups which are associated with a differential prognosis (26). Since it has been shown that patients with a good prognosis can be cured with less intensive treatments than patients with a poor prognosis, a risk-stratified treatment protocol is currently applied to treat patients with this disease (21,26,189). Evidence about the prognosis of newly identified subgroups resulted in a reclassification of patients to other risk groups. According to the reclassification, two types of personalized medicine strategies can be identified at this moment. Patients reclassified to the favorable risk group receive a less intensive treatment, while a more intensive treatment is administered to patients identified with unfavorable risk group. More information about the personalized medicine strategies can be found in the supplementary material. The cost-effectiveness of the two types of personalized medicine strategies are separately evaluated in this study. Both analyses are restricted to patients aged 18-60 years because treatment only differs between risk groups for these patients.

Decision model

A validated decision model (243) was used to perform the cost-effectiveness analyses of the personalized medicine strategies. The model is an individual patient simulation in which transitions between health states were based upon estimated time to (second) CR, relapse and death. These estimates were derived from patient-level data (189) and

differed between relevant patient and disease characteristics including risk classification.

Input parameters

The cost-effectiveness analyses were performed from a Dutch health-care perspective. All input parameters were derived from previous studies (Table S8.1 in the supplementary material). Treatment effects were included by the hazard ratios (HR) for treatment-related mortality (TRM), relapse risk and survival after relapse. The HRs were either directly taken from the literature or estimated according to the method of Parmar et al. (244). A random-effects meta-analysis was performed if more than one HR was available. Costs were categorized as diagnostic, induction, post-remission, follow-up and relapse costs. Follow-up costs were distinguished in standard follow-up visits for regular visits to test for leukemia recurrence and complication costs. The complication costs included costs of additional visits, hospital stay and medication. All costs were converted to 2013 Euros. Quality of life utilities were derived from EQ-5D estimates in AML patients aged 18-60 years. The estimates consisted of a baseline quality of life utility and decrements for patients with relapse/refractory AML and patients treated with an allogeneic hematopoietic stem cell transplantation.

Sensitivity analyses

Probabilistic sensitivity analyses (PSA) with 1,000 simulations were performed to assess the impact of uncertainty in the input parameters on the cost-effectiveness. Within each simulation, the values of the input parameters were randomly drawn from distributions shown in Table S8.2 of the Supplementary Material. Acceptability curves were constructed to identify the likelihood that the personalized medicine strategies were cost-effective given the uncertainty in the input parameters.

Evaluation of the required methodology for cost-effectiveness analyses of personalized medicine strategies

The cost-effectiveness analyses were first restricted to an assessment of the consequences of a treatment change in the newly identified subgroups (defined as treatment-only approach). Subsequently, the consequences of the test were added to the analyses (defined as test-treatment approach). It was assumed that the tests had no direct effect on health, but only impacted health by means of a treatment change. The formulas in Box 8.1 were used to calculate the incremental cost-effectiveness ratio (ICER) of the treatment and test-treatment approach.

For the sake of the illustration, it was assumed that perfect tests (100% sensitivity and specificity) were used to identify the new subgroups in AML. However, a reduced per-

formance of the test is easy to include in the analysis. Assuming that a positive test is associated with a treatment change, it is required to assess the health and economic consequences of that treatment change separately for true and false positive tested patients. The overall incremental costs and effects is estimated a weighted average of the consequences in the two patient groups (formula 2.1 in Box 8.1). Since treatment does not change for negatively tested patients, a lower specificity impacts the cost-effectiveness analyses of personalized medicine strategy by reducing the size of the subgroup with a treatment change (formula 2.2 in Box 8.1).

Box 8.1 Formulas initially used to estimate the cost-effectiveness of personalized medicine strategies.

$$\text{ICER treatment-only approach} = \frac{\Delta \text{ Costs (treatment change)}}{\Delta \text{ Effectiveness (treatment change)}} \quad (1)$$

$$\text{ICER test-treatment approach} = \frac{\text{Costs of the test} + p^* \Delta \text{ Costs (treatment change)}}{p * \Delta \text{ Effectiveness (treatment change)}} \quad (2)$$

$$\Delta \text{ Costs (treatment change)} = \frac{TP}{TP+FP} * \Delta \text{ Costs}_{TP} + \frac{FP}{TP+FP} * \Delta \text{ Costs}_{FP} \quad (2.1)^1$$

$$\Delta \text{ Effectiveness (treatment change)} = \frac{TP}{TP+FP} * \Delta \text{ Effectiveness}_{TP} + \frac{TP}{TP+FP} * \Delta \text{ Effectiveness}_{FP} \quad (2.2)^1$$

$$p = \frac{TP+FP}{TP+FP+TN+FN} \quad (2.3)$$

[1] These formulas assume that treatment changes for patients with a positive test results. If otherwise, the costs and effects are weighted according to the proportion true and false negative patients.

Δ Costs treatment change = Differences of all health care costs, except costs of testing, between the new and old treatment strategy

Δ Effectiveness treatment change = differences in both life expectancy and quality-adjusted life years (QALYs) between the new and old treatment strategy

p = proportion of patients in the total tested population whose treatment changes

TP = proportion of patients in the total tested population with a true positive test result (correctly classified as mutant)

FP = proportion of patients in the total tested population with a false positive test result (incorrectly classified as mutant)

TN = proportion of patients in the total tested population with a true negative test result (correctly classified as not mutant)

FN = proportion of patients in the total tested population with a false negative test result (incorrectly classified

The impact of the exclusion of the test was assessed by comparing the base-case results and acceptability curves of the treatment-only and test-treatment approach. Within this evaluations, four generic parameters in cost-effectiveness analyses of personalized medicine strategies were varied to assess the consequences of these parameters on the difference in ICER between the two approaches. These four parameters were 1) incremental costs of treatment change, 2) incremental health effects of treatment change, 3) prevalence of the subgroup with a treatment change and 4) testing costs.

RESULTS

Case study results

Table 8.1 shows the results of the cost-effectiveness analyses of the two personalized medicine strategies. The less intensive treatment for favorable risk patients resulted in lower medical costs (€152,536 and €158,154 for the new and current protocol, respectively). Health was improved as indicated by the increased life years and QALYs (health gain is 0.851 life years and 1.005 QALYs). The more intensive treatment for unfavorable risk patients was associated with larger medical costs (€246,838 versus €168,898), improved life expectancy (5.48 versus 4.56 years) and increased QALYs (4.25 versus 3.65).

The incremental effects decreased substantially in the test-treatment approach, because health gains were only observed for the small group of patients with a treatment change (Table 8.1). The observed health gain in patients with a treatment change did not differ from the treatment-only approach. The impact of the inclusion of the test on the incremental costs differed between the personalized medicine strategies for the newly favorable and unfavorable risk patients. Since both patients with and without a treatment change need to be tested, the incremental costs for all patients increase with the costs of testing. The impact on the overall incremental costs depends on the costs of the test relatively to the incremental costs of a treatment change. The costs of the test were larger than the incremental costs of a treatment change in newly favorable patients. Consequently, an increase in the overall incremental costs was observed if the costs of testing were included in the analysis. Contrary, a decrease in incremental costs was observed for the strategy in newly unfavorable patients, because the costs of testing were smaller than the incremental costs of the treatment change for these patients. Despite the differential effect on the incremental costs in the two personalized medicine strategies, the ICER increased for both strategies if the costs of testing were included in the analysis (Table 8.1).

Table 8.1 Cost-effectiveness results in the base-case analysis

	%	Costs (€)			Life years			QALYs			ICER (Costs/LY)	ICER (Costs/QALY)
		Current strategy	New (PM-) strategy	Difference	Current strategy	New (PM-) strategy	Difference	Current strategy	New (PM-) strategy	Difference		
New favorable risk group												
Treatment-only approach	100.0	158,184	152,536	-5,649	20.69	21.54	0.851	17.29	18.30	1.005	-6,639	-5,623
Test-treatment approach												
- Treatment change	9.0	158,184	153,167	-5,017	20.69	21.54	0.851	17.29	18.30	1.005	-5,897	-4,994
- No treatment change	91.0	TC	TC + 631	631	LE	= LE	0.000	QALE	= QALE	0.000	NA	NA
Weighted average				124			0.076			0.090	1,623	1,374
New unfavorable risk group												
Treatment-only approach	100.0	168,898	246,838	77,939	4.56	5.48	0.925	3.65	4.25	0.599	84,282	130,065
Test-treatment approach												
- Treatment change	0.4	168,898	247,048	7,815	4.56	5.48	0.925	3.65	4.25	0.599	84,510	130,417
- No treatment change	99.7	TC	TC + 210	210	LE	= LE	0.000	QALE	= QALE	0.000	NA	NA
Weighted average				483			0.003			0.002	149,289	230,386

PM = personalized medicine, TC = treatment costs, LE = life expectancy, QALE = Quality-adjusted life expectancy, QALYs = quality-adjusted life years, ICER = incremental cost-effectiveness ratio, NA = not applicable (because incremental effects = 0). The costs and effects estimates of the 'no treatment change' group were not estimated in this study as the actual values will not influence the incremental costs and effects. Both the costs of treatment as the health effects will be similar in the treatment alternatives, because treatment did not change for this group of patients.

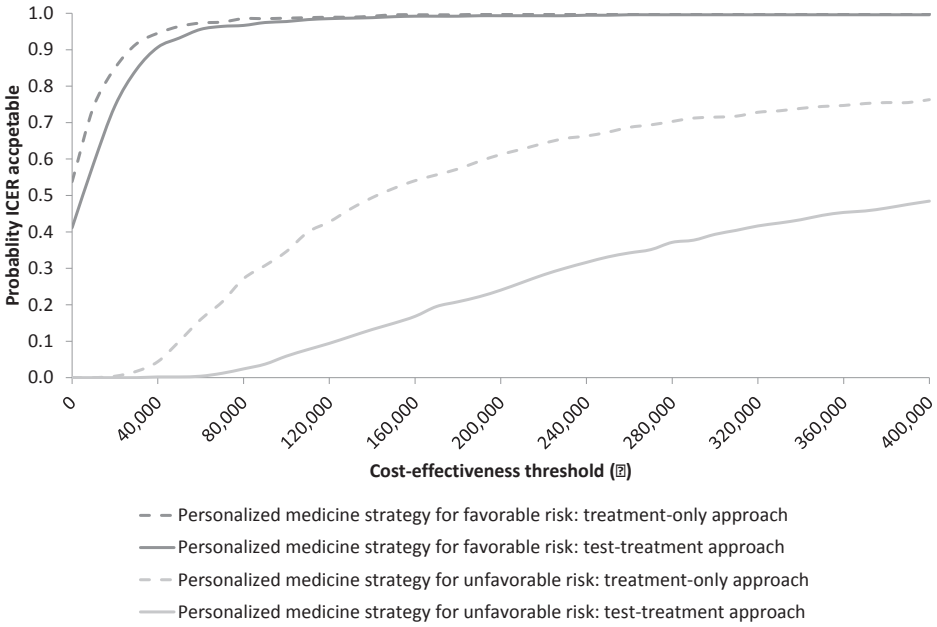


Figure 8.1 Acceptability curves of the comparison of new personalized medicine strategies versus current strategies

The probabilistic sensitivity analyses identified the probability that the new personalized medicine strategies are cost-effective given the uncertainty in the input parameters and different threshold values. These probabilities were reported in acceptability curves as shown in Figure 8.1 Assuming a threshold of €80,000 per QALY, the probability of being cost-effective is 97% and 2% for the personalized medicine strategies (including costs of testing) in newly favorable and unfavorable risk patients, respectively. These probabilities are higher if the analyses is restricted to the treatment-only approach. Consequently, the impact of uncertainty in the input parameters is underestimated if analyses are restricted to the treatment-only approach.

Relationship between the treatment-only and test-treatment approach

The treatment-only approach systematically underestimates the ICER due to the exclusion of the testing costs. The severity of the underestimation depends on the costs of the test, the incremental health effects of the treatment change and the prevalence of the subgroup with a treatment change (Figure 8.2). The difference in the ICER between the test-treatment and treatment-only approach increases linearly with higher costs of testing. Furthermore, the difference in the ICER decreases with a diminishing rate if the prevalence of the subgroup with a treatment change increases or if the incremental health effects of the treatment change increases. Although the ICERs of both the

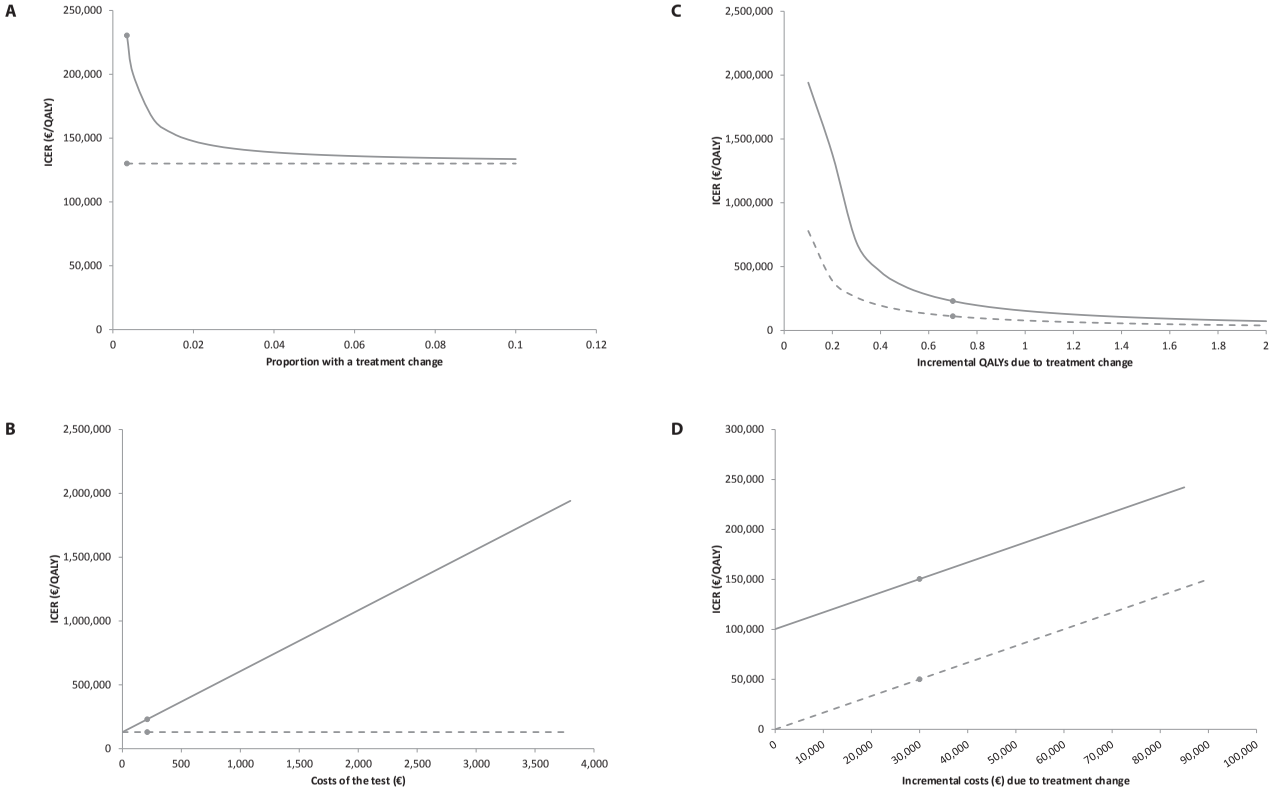


Figure 8.2 The impact of four main parameters (A = proportion with a treatment change, B = costs of the test, C = incremental QALYs due to treatment change, and D = incremental costs due to treatment change) on the cost-effectiveness of the newly unfavorable risk group.

The dashed line shows the ICER of the treatment-only approach and the solid line shows the ICER of the test-treatment approach. The dots show the results of the base-case analysis. ICER=incremental cost-effectiveness ratio, QALY=quality-adjusted life year

treatment-only and test-treatment approach increase with larger incremental costs due to the treatment change, the difference between the two ICERs remains constant. These findings regarding the relationship between the ICER of the treatment change and the impact of the test on the cost-effectiveness is expressed in a newly developed formula (formula 3, see Box 8.2).

Box 8.2 Formula and illustration of a method to estimate the cost-effectiveness of personalized medicine strategies

ICER test-treatment approach = ICER treatment-only approach + impact test approach (3)

$$\frac{\Delta \text{ costs (treatment change)}}{\Delta \text{ effectiveness (treatment change)}} + \frac{\text{Costs of test}}{\Delta \text{ effectiveness (treatment change)}} * \frac{1}{p}$$

Illustration of formula:
Favorable subgroup

ICER of treatment-only approach $= \frac{-5{,}649}{1.005} = -5{,}621$ per QALY gained

Impact of test approach $= \frac{631}{1.005} * \frac{1}{0.09} = 6{,}976$ per QALY gained

ICER of test-treatment approach $= -5{,}621 + 6{,}976 = 1{,}355$ per QALY gained

Unfavorable subgroup

ICER of treatment-only approach $= \frac{77{,}939}{0.599} = 130{,}115$ per QALY gained

Impact of test approach $= \frac{210}{0.599} * \frac{1}{0.004} = 87{,}646$ per QALY gained

ICER of test-treatment approach $= 130{,}115 + 87{,}646 = 217{,}761$ per QALY gained

(All numbers are derived from table 1, differences in the ICER are due to rounding))

DISCUSSION

The aim of this study was to evaluate different methods for cost-effectiveness analyses of personalized medicine strategies in order to develop more specific methodological recommendations for these analyses. The use of one standardized methodology in future analyses may improve the comparability of study results and enable better judgments regarding the added value of new personalized medicine strategies.

This study showed that the cost-effectiveness is overestimated if the costs of testing are excluded from the analysis. The overestimation is especially problematic for personalized medicine strategies with large costs of testing, a small prevalence of the subgroup with a treatment change and small health effects of that treatment change. In addition, it was shown that the impact of uncertainty in the input parameters was underestimated in analyses restricted to the subgroup with a treatment change. According to these findings, it is strictly recommended to include the costs of testing in cost-effectiveness analyses of personalized medicine strategies. Nonetheless, the primary analyses should always be focused on the cost-effectiveness of the treatment change in the selected subgroups, because a cost-effective treatment is a prerequisite for a cost-effective personalized medicine strategy. Furthermore, it provides insight in the achieved health gains for the selected subgroup. We developed a formula to incorporate the test in the cost-effectiveness analysis of personalized medicine strategies without losing any information regarding the health and economic consequences of the treatment change.

The formula was tested in two personalized medicine strategies for patients with acute myeloid leukemia. However, it is expected that it also applies to other disease areas, because only generic components of personalized medicine strategies were included in the formula. Furthermore, several other studies indicated these components as driving factors in the cost-effectiveness of personalized medicine strategies (245-249).

Reimbursement agencies need to be aware of the systematic bias in studies restricted to cost-effectiveness of new targeted treatments without including the costs of testing. These studies are only allowed if the identification of the subgroup is not associated with additional costs because the test is already part of standard clinical practice. The developed formula is also useful for reimbursement agencies in the interpretation of cost-effectiveness studies and selection of the most appropriate reimbursement strategy. For example, in the circumstance that the treatment change in the selected subgroup is cost-effective, but the overall personalized medicine strategy not, it might be worthwhile to negotiate a price reduction of the test or limit the use of the test to allow access to the new treatment. Negative reimbursement decisions would be more likely if the negotiations about the costs of the treatment are especially required if the treatment change is not cost-effective.

The use of the formula provides also advantages for the analysis and reporting of cost-effectiveness analyses of personalized medicine strategies. Health outcomes and costs only need to be collected and analyzed for the subgroup whose treatment changes. This restricted data collection and analysis limits the required time and costs and allows faster implementation. Secondly, the formula can easily be applied early in the develop-

ment process of personalized medicine strategies in decisions about the continuation of the development. It may identify areas for improvement in both the test and treatment strategy and assess whether it is worthwhile to continue the development given the expected cost-effectiveness. These early motivated decisions may result in a better allocation of resource & development costs (41). Another advantage of the early use of the formula in the development process is that these analyses may identify critical parameters in the cost-effectiveness analysis for which additional data is required. Adequate collection of that data during the development of the personalized medicine strategy may increase the available evidence at time of market entrance and thereby improves the implementation of personalized medicine strategies (250).

SUPPLEMENTARY MATERIAL

The case study evaluated the cost-effectiveness of changes in the risk-stratified treatment of AML in patients aged 18-60 years. For many years, risk-classification of AML was solely based upon cytogenetic abnormalities (251). Recent studies have shown that the risk classification could be improved by including molecular abnormalities.(252) Patients with intermediate cytogenetics and *CCAAT enhancer binding protein* double mutations (*CEBPAdm*) or *nucleophosmin-1 gene mutations* (*NPM1)* without internal tandem duplication of the *fms-like tyrosine kinase-3 gene* (*FLT3*-ITD) have a favorable prognosis,(191,253-257) while patients with intermediate cytogenetics and *ecotropic virus integration-1* (*EVI1)* overexpression have an unfavorable prognosis.(258) This new information results in treatment changes because the choice of post-remission treatment differs between risk groups.(26) High-dose chemotherapy is the only post-remission treatment in favorable risk patients. An allogeneic hematopoietic stem cell transplantation (HSCT) from a sibling donor is the preferred post-remission treatment option in all other patients. If no sibling donor is available, intermediate risk patients will receive an autologous HSCT or high-dose chemotherapy, while unfavorable risk patients may receive an allogeneic HSCT from a matched unrelated donor or an umbilical cord blood transplantation. In general, the better the prognosis, the least intensive treatment is administered. Consequently, patients with *NPM1* mutations without *FLT3*-ITD or *CEBPA*dm (newly favorable) will no longer receive an allogeneic or autologous HSCT and patients with *EVI1* overexpression (newly unfavorable) become candidates for an umbilical cord blood transplantation or allogeneic HSCT from a matched unrelated donor.

Figure S8.1 shows the consequences in treatment if all AML patients aged 18-60 years are tested for the new abnormalities. As only patients with *NPM1* mutations without *FLT3*-ITD *CEBPA*dm or *EVI*1 overexpression will be reclassified to another risk group, treatment will not change for all other patients. Furthermore, within the newly identified subgroup, treatment does also not change for patients without a complete remission because only post-remission treatment differs between risk groups. Finally, even a selective group of patients with a complete remission will receive a different treatment because stem cell transplantations are not feasible for all patients. Consequently, it was found that 9% and 0.4% of all AML patients aged 18-60 received a different treatments due to the reclassification to the favorable or unfavorable risk group, respectively.

Table S 8.1 Values and distributions of the input parameters

	Base-case	Probabilitic sensitivity analysis		Source
		Distribution	Parameters	
Effectiveness post-remission treatment				
HR TRM: allogeneic HSCT versus chemotherapy	3.60	Lognormal	μ=1.28, σ=0.39	Schlenk et al. 2008(259)
HR TRM: autologous HSCT versus chemotherapy	2.58	Lognormal	μ=0.95, σ=0.28	Nathan et al. 2004(260), Breems et al. 2005(261), Vellenga et al. 2011(205)
HR TRM: allogeneic HSCT versus autologous HSCT	2.18	Lognormal	μ=0.78, σ=0.28	Ringden et al. 2000(262), Suciu et al. 2003(263), Brunet et al. 2004(264), Lazarus et al. 2006(265)
HR relapse: allogeneic HSCT versus chemotherapy	0.49	Lognormal	μ=-0.72, σ=0.29	Schlenk et al. 2008(259)
HR relapse: autologous HSCT versus chemotherapy	0.84	Lognormal	μ=-0.42, σ=0.22	Breems et al. 2005(261), Vellenga et al. 2011(205)
HR relapse: allogeneic HSCT versus autologous HSCT	0.66	Lognormal	μ=-0.18, σ=0.27	Ringden et al. 2000(262), Suciu et al. 2003(263), Lazarus et al. 2006(265)
HR death after relapse: HSCT versus no HSCT	1.35	Lognormal	μ=0.30, σ=0.10	Breems et al. 2005(207)
Costs				
- Costs chemotherapy	34,225	Gamma	α=13.4, β=2,562.5	Leunis et al. 2013(266)
- costs autologous HSCT	33,277	Gamma	α=10.1, β=3,286.7	Leunis et al. 2013(266)
- Costs allogeneic HSCT from sibling donor	44,070	Gamma	α=2.4, β=18,032.3	Leunis et al. 2013(266)
- Costs allogeneic HSCT from MUD	82,041	Gamma	α=26.9, β=3,052.0	Leunis et al. 2013(266)
- Costs UCB transplantation	109,675	Gamma	α=5.9, β=18,466.9	Blommestein et al. 2012(71)
- Costs complications chemotherapy	1,909	Gamma	α=0.09, β=22,302	Leunis et al. 2013(266)
- Costs complications autologous HSCT	1,942	Gamma	α=0.06, β=32,955	Leunis et al. 2013(266)
- Costs complications allogeneic HSCT from sibling donor	18,061	Gamma	α=0.56, β=32,404	Leunis et al. 2013(266)
- Costs complications allogeneic HSCT from MUD	36,521	Gamma	α=1.02, β=35,832	Leunis et al. 2013(266)
- Costs complications UCB transplantation	61,230	Gamma	α=0.73, β=83,495	Blommestein et al. 2013(71)
- Costs complication induction treatment	7,793	Gamma	α=0.17, β=45,653	Leunis et al. 2013(266)

Table S 8.1 Values and distributions of the input parameters (continued)

	Base-case	Probabilitic sensitivity analysis		Source
		Distribution	Parameters	
- Costs standard follow-up visit	253	Uniform	Min=126, Max=379	Standard Dutch tariffs
- Costs bone marrow aspirate	306	Uniform	Min=153, Max=458	Standard Dutch tariffs
- Reinduction chemotherapy favorable risk	45,610	Gamma	α=8.6, β=5,320.2	Leunis et al. 2013(266)
- Relapse treatment intermediate or unfavorable risk	43,461	Gamma	α=8.6, β=5,320	Uyl-de Groot et al. 2001(63)
- Assumption follow-up costs after 1 year	0.25	Uniform	Min=0, Max=1	van Agthoven et al. 2002(267)
Quality of life utility				
- Utility no allogeneic HSCT + no relapse/refractory disease	0.86	Beta	α=137, β=22	Leunis et al. 2014(172)
- Decrement allogeneic HSCT	0.03	Beta	α=1, β=24	Leunis et al. 2014(172)
- Decrement relapse/ refractory treatment	0.09	Beta	α=4, β=35	Leunis et al. 2014(172)
Risk group distribution				
Proportion favorable cytogenetics	0.15		α_1=62	Leunis et al. 2013(243)
Proportion intermediate risk group	0.69	Dirichlet	α_2=293	Leunis et al. 2013(243)
Proportion unfavorable cytogenetics	0.17		α_3=72	Leunis et al. 2013(243)
Proportion normal karyotype within intermediate risk group	0.71	Beta	α=199, β=81	Derived from data used in Leunis et al. 2013(243)
Proportion CEBPAdm within normal karyotype	0.06	Beta	α=139, β=2104	Dufour et al. 2010(253), Green et al. 2010(254), Schlenk et al. 2008(190)
Proportion NPM1+/FLT3- within normal karyotype	0.27	Beta	α=396, β=1052	Thiede et al. 2006(257), Boissel et al. 2005(255), Döhner et al. 2005(256), Schneider et al. 2009(268)
Proportion EVI1 overexpression within normal karyotype	0.05	Beta	α=31, β=559	Gröschel et al. 2010(258)

Table S 8.1 Values and distributions of the input parameters (continued)

	Base-case	Probabilitic sensitivity analysis		Source
		Distribution	Parameters	
CR rates new mutations				
CR rate CEBPAdm	0.91	Beta	α=137, β=14	Dufour et al. 2010(253), Green et al. 2010(254), Schlenk et al. 2008(190)
CR rate NPM1+/FLT3-	0.77	Beta	α=297, β=90	Thiede et al. 2006(257), Döhner et al. 2005(256), Schneider et al. 2009(268)
CR rate EVI1 high expression	0.71	Beta	α=44, β=18	Gröschel et al. 2010(258)
Treatment distribution				
Chemotherapy favorable risk group	0.92	Beta	α=55, β=5	Leunis et al. 2013(243)
No post-remission treatment intermediate risk group	0.13		α_1=29	Leunis et al. 2013(243)
Chemotherapy intermediate risk group	0.19	Dirichlet	α_2=43	Leunis et al. 2013(243)
Autologous HSCT intermediate risk group	0.32		α_3=74	Leunis et al. 2013(243)
Allogeneic HSCT intermediate risk group	0.36		α_4=82	Leunis et al. 2013(243)
No post-remission treatment unfavorable risk group	0.11		α_1=8	Leunis et al. 2013(243)
Chemotherapy unfavorable risk group	0.17		α_2=13	Leunis et al. 2013(243)
Autologous HSCT unfavorable risk group	0.12	Dirichlet	α_3=9	Leunis et al. 2013(243)
Allogeneic HSCT sibling donor unfavorable risk group	0.36		α_4=27	Leunis et al. 2013(243)
Allogeneic HSCT MUD unfavorable risk group	0.23		α_5=17	Leunis et al. 2013(243)
CB transplantation MUD unfavorable risk group	0.01		α_6=2	Leunis et al. 2013(243)
Costs				
- Costs new diagnostic tests newly favorable	641	Uniform	Min=666, Max=1999	Standard Dutch tariffs
- Costs new diagnostic test newly unfavorable	240	Uniform	Min=120, Max=360	Standard Dutch tariffs

HSCT=hematopoietic stem cell transplantation, MUD=matched unrelated donor, CR=complete remission, UCB=umbilical cord blood, HR=hazard ratio, TRM=treatment-related mortality

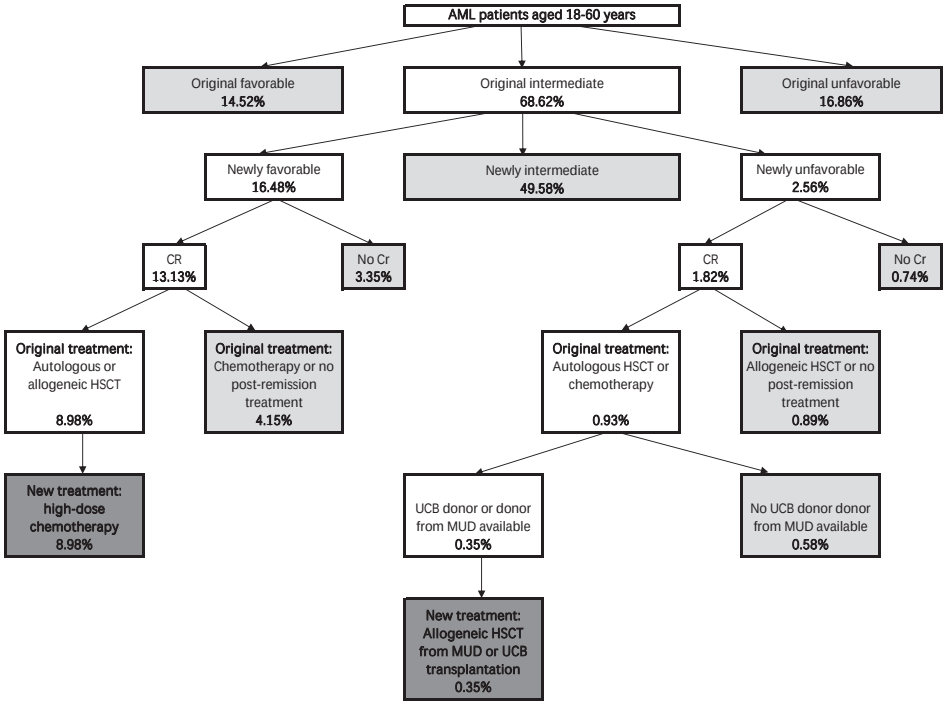
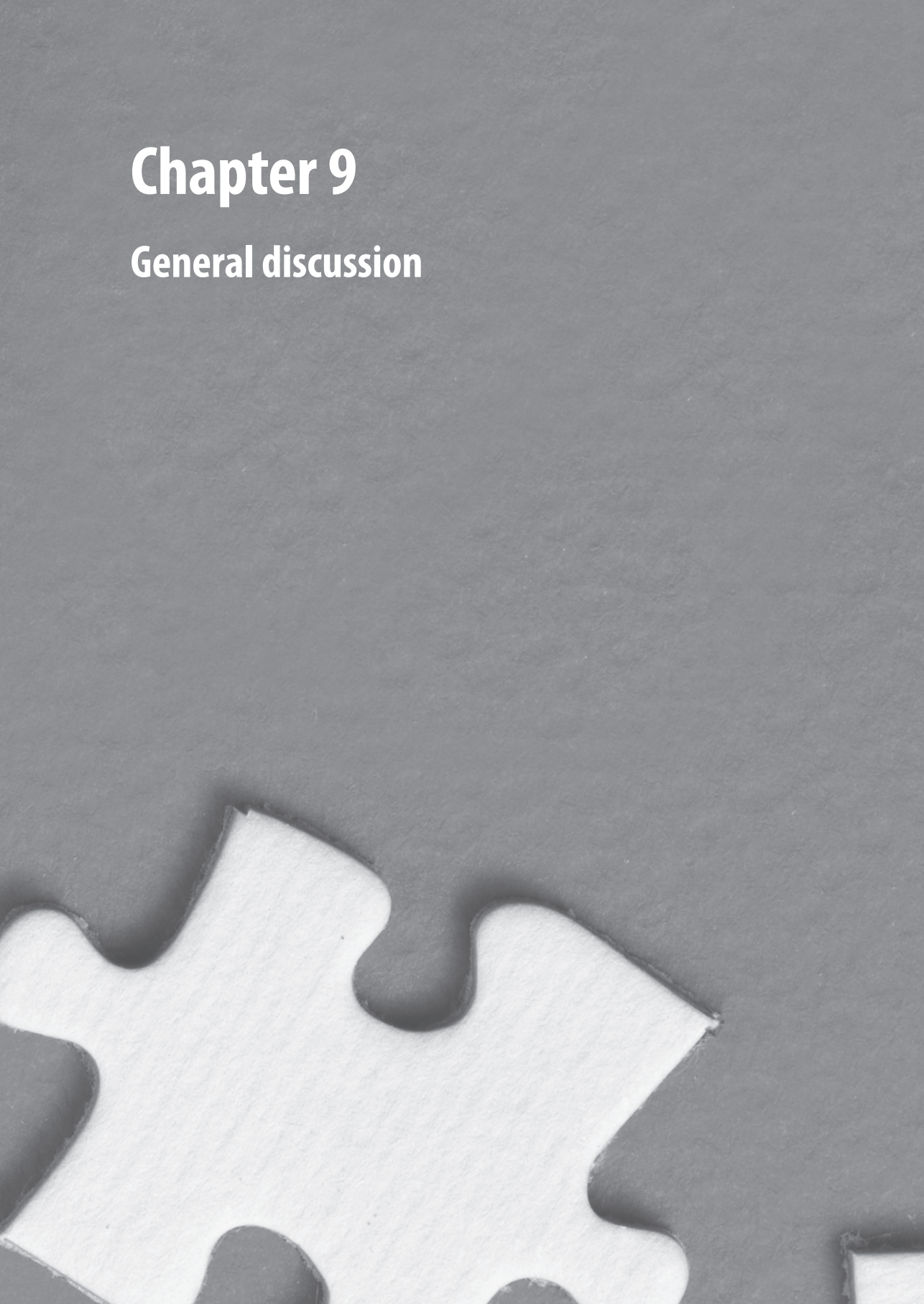


Figure S8.1 Patient flow of AML patients.

The light-gray blocks indicate no treatment change. The dark-grey blocks indicate treatment change. Original favorable = patients with t(8;21) or inv(16) abnormalities. Original unfavorable = patients with complex karyotype (≥3 cytogenetic abnormalities), inv(3), t(6;9), -5/del(5q), -7/del(7q), abn(11q23) or t(9;22). Original intermediate = all other patients. Newly favorable = patients with normal karyotype and CEBPAdm or NPM1+ without FLT3ITD. Newly unfavorable risk group = patients with normal karyotype and EVI1 overexpression. Newly intermediate = all other original intermediate patients. CR = complete remission. HSCT = hematopoietic stem cell transplantation, MUD = matched unrelated donor, UCB = umbilical cord blood donor

Chapter 9

General discussion



INTRODUCTION

This chapter describes the main findings in this thesis and indicates areas for future research by answering three questions: 1) Which areas for future improvements in the field of acute myeloid leukemia have been identified in this thesis, 2) Which elements are essential in cost-effectiveness analyses of personalized medicine strategies, and 3) Which other aspects should be considered to improve the implementation of personalized medicine strategies?

AREAS FOR FUTURE IMPROVEMENT IN THE FIELD OF ACUTE MYELOID LEUKEMIA (AML)

Adequate treatment for AML has been a challenge for many decades. The chemotherapy regimen has only marginally changed since 1960. However, the introduction of hematopoietic stem cell transplantation (HSCT) improved the cure rate of AML (120), but at the cost of more severe side effects and higher treatment costs. This thesis showed substantially greater costs for allogeneic stem cell transplantation compared to chemotherapy due to a longer hospital stay and the costs of donor searching and treating the post-transplant complication graft-versus-host-disease. Furthermore, quality of life seems to be worse in patients who have received a stem cell transplantation. Since the large negative consequences of allogeneic stem cell transplantation are commonly known in the field, these procedures are restricted to patients who are less likely to be cured with non-transplant therapeutic approaches (26). Patients with a favorable prognosis are initially treated with high-dose chemotherapy and only recommended for a stem cell transplantation in case of leukemia recurrence (269).

Recent research in the field of AML focuses on the further improvement of treatment outcome by new drug development and better stratification of patients for selecting therapies. The stratification is based upon both clinical factors and underlying cytogenetic and molecular abnormalities of the disease. Evidence about the prognostic impact of newly identified molecular subgroups, such as AML's with *nucleophosmin-1 gene mutations* (*NPM1)* without internal tandem duplication of the *fms-like tyrosine kinase-3 gene* (*FLT3*-ITD)(255-257), *CCAAT enhancer binding protein gene* double mutations (*CEBPA*dm) (253,254) and overexpression of *ecotropic virus integration-1 transcript* (*EVI1*) (258), resulted in a reclassification of these patients to either the favorable or unfavorable risk group. Since the treatment options differs between risk groups, the reclassification causes a treatment change for the patients in the newly identified subgroups.

This thesis showed that the treatment change due to the reclassification of patients towards the favorable risk group (patients with *CEBPA*dm or mutated *NPM1* without *FLT3*-ITD) improves life expectancy and quality of life. The treatment-related mortality is lower for high-dose chemotherapy and fewer treatment-related complications arise. Furthermore, since high-dose chemotherapy is less expensive than allogeneic HSCT, treatment costs are reduced. However, the total costs of the improved stratification increased due to the additional costs of molecular testing. Nevertheless, the improved stratification of patients with *CEBPA*dm or mutated *NPM1* without *FLT3*-ITD can be considered cost-effective given an incremental cost-effectiveness ratio (ICER) of €1,374 per quality-adjusted life year (QALY).

Due to the reclassification of AML patients with *EVI1* overexpression to the unfavorable risk group, these patients became candidates for an allogeneic HSCT from a matched unrelated donor or umbilical cord blood transplantation if a suitable sibling donor was unavailable. These treatments improved the survival of patients with *EVI1* overexpression, but the additional life years were spent with a reduced quality of life due to complications of the treatment (graft-versus-host-disease). Furthermore, the treatment costs were much higher than the costs of the alternative treatments (chemotherapy or autologous HSCT). This treatment change costs €130,065 per QALY gained. It can therefore be argued that it is not cost-effective to treat patients with *EVI1* overexpression with an allogeneic HSCT from a matched unrelated donor or cord blood transplantation. For these patients, more effective and/or less toxic treatments are needed. Currently, research is ongoing to develop such treatments by searching for targeted drugs.

Even if a cost-effective treatment option for patients with *EVI1* would be available in the future, a personalized medicine strategy to identify and better treat this subgroup is not by definition cost-effective. Chapter 8 showed that the inclusion of the testing costs always increases the ICER. The rate of the increase depends on the costs of the test and the size of the subgroup with a treatment change relative to the total tested population. The increase is especially substantial for small subgroups. As *EVI1* overexpression only occurs in 3% of all AML patients, it may be challenging to develop cost-effective personalized medicine strategies for these patients. However, new techniques like gene expression profiling and whole genome sequencing might minimize this disadvantage as these tests enable the identification of more than one molecular aberration in one test. Consequently, the identification of *EVI1* overexpression would not require any additional costs of testing.

The use of gene expression profiling and whole genome sequencing also enable the identification of other (currently unknown) therapeutically relevant subgroups. These

new subgroups might lead to the identification of new drug targets for future treatment development in AML. At this moment, new agents are already in development that target leukemias with *FLT3* and *IDH* (*isocitrate dehydrogenase*) gene mutations (270). It is important that the studies that evaluate new treatment modalities in AML not only focus on improved survival, but also evaluate the impact on quality of life. For many years, the emphasis has been on improving survival for AML patients. This focus was legitimate in that period because AML had a large fatality rate. However, since the cure rate has improved over the last decades (119,120), greater attention should be paid to quality of life since it has become apparent that survivors of AML experience many problems with functioning in daily life. As targeted treatments have a specific mode of action, it is expected that these treatments in many instances may have fewer side effects and thereby allow for improved quality of life than current treatments.

Since different questionnaires are available to measure quality of life, it is difficult for investigators to choose the most appropriate method. Furthermore, the aim of quality of life assessment may differ between clinicians and others like health economists and policymakers. From a clinical perspective, a more detailed and disease-specific questionnaire, like the EORTC quality of life questionnaire for cancer (QLQ-C30), might be desirable since it provides specific information about clinically relevant problems (29). In contrast, a more generic questionnaire, like the EuroQol-5D (EQ-5D), is preferred from a societal perspective as these questionnaires enable the comparison of quality of life across diseases (161). In order to converge the different aims in one questionnaire, this thesis derived quality of life utilities for different disease- specific questionnaires, including the QLQ-C30, by mapping and direct valuation. It was shown that the average utilities derived from the mapping algorithms were comparable with EQ-5D utilities. However, items were only included in the mapping algorithm if these were also captured by the EQ-5D. This means that certain disease-specific elements are still excluded in the utility measurement by means of mapping. These disease-specific elements are better included by the direct valuation methods. Nevertheless, the utilities from the disease-specific preference based questionnaire were significantly higher with a smaller range compared to the EQ-5D utilities. This smaller ranges mean that the total possible quality of life gain is smaller for disease-specific utilities. Consequently, the use of utilities from disease-specific preference-based instruments in cost-effectiveness analyses may limit the comparability of cost-effectiveness results across diseases. It is therefore still recommended to include both generic and disease-specific questionnaires when assessing quality of life. Fortunately, since the EQ-5D is short, it is feasible to use both in the same study.

The inclusion of both generic and disease-specific questionnaires in future studies will also enable a further assessment of the psychometric properties of these two questionnaires. That information is needed to provide additional information about the validity of the EQ-5D in acute leukemia patients. Although, the EQ-5D is the preferred quality of life questionnaire for economic evaluations for reasons of comparability, reimbursement authorities like NICE allow for the use of other measures provided that the EQ-5D is not valid in that specific population (161), for example in mental health (271). Chapter 6 of this thesis assessed the validity of the EQ-5D in AML survivors and did not find any significant problems regarding the validity. However, more evidence is needed since only a selective group of AML patients was included, repetitive measures of quality of life were unavailable and no objective measure of disease status was included in the study. These shortcomings can be overcome if quality of life will be measured alongside prospective clinical trials.

The quality of life of AML can also be further improved by the use of adequate supportive care programs. Clinicians should identify which patients need supportive care and select the most appropriate program for each patient. The best approach can be selected by using patient-reported outcomes measures, such as the QLQ-C30, to identify the patient's quality of life problems. Furthermore, individual patient characteristics such as age, ethnicity, social economic status and marital status may also influence the effectiveness of different supportive care programs (272,273). At this moment, only a few supportive care programs have been evaluated in AML survivors (274). Although more evidence is available from other cancer survivors (275), these findings need to be confirmed in AML survivors.

Another important patient-reported outcome parameter is patient satisfaction. Satisfaction might be improved by reducing the hospital stay during active treatment of AML. Nowadays, patients are often hospitalized for about one month because they are at high risk of infections. Nevertheless, outpatient treatment policies may be feasible in at least some AML patients (80,81). Further research is needed regarding these outpatient treatment options. A reduction in hospital stay will almost certainly not only improve patient satisfaction, but also reduce the treatment costs. It might be possible that new targeted treatment can be more easily applied in an outpatient setting, because it is conceivable that some of these treatments will be associated with reduced toxicities and medical complications. Thereby, the added value of the new targeted treatments might not only be found in clinical outcomes such as survival and quality of life, but also in reduced costs. Cost-effectiveness of new technologies and therapeutic compounds are needed to show this added value and receive adequate reimbursement for these new products.

Adequate reimbursement is an important prerequisite for successful implementation of new products in clinical practice. Hospitals in the Netherlands and other countries experience more and more financial risk due to the introduction of managed competition in the health care sector. Consequently, hospitals need to receive sufficient reimbursement for the treatments to compensate the costs of treating patients and remain financially viable. Transparency about the actual costs of technologies and treatment is necessary for adequate reimbursement is only feasible if information is available about the actual costs of technologies and treatments. Findings from the past and present show that adequate reimbursement is not only difficult for new treatments but also for existing treatments. In 2009, hospitals faced losses regarding the treatment of acute leukemia and initiated a costing study to demonstrate the actual treatment costs of acute leukemia. The results of this study are reported in chapter 2. It was shown that the received reimbursement of one chemotherapy cycle was only 25% of the actual costs (€10,000 instead of €40,000) (276). This considerable difference was the automatic consequence of the use of standard tariffs for a hospital stay. However, the care for AML patients requires highly specialized treatment care and equipment and a hospital day is therefore much more costly than a hospital day for patients who, for instance, undergo a knee replacement (69,277). Furthermore, the observed hospital stay during the chemotherapy cycle was much longer than assumed in the reimbursement tariff. A discussion of the findings with the Dutch Health Care authority improved the reimbursement of chemotherapy cycles. The reimbursement of stem cell transplantations also changed according to costs studies. However, a dramatic drop in the reimbursements of the post-transplantation trajectory was observed in 2014 (276). Research is ongoing with respect to the reasons and consequences for this discrepancy.

In the Netherlands and several other countries, it is recommended to perform cost-effectiveness analysis from a societal perspective (278). Therefore, future studies should not only measure the direct medical costs, such as the costs of medication, hospital visits and laboratory tests, but also the costs of productivity loss due to work absenteeism and informal care costs. Chapter 5 showed that a large proportion of surviving patients with AML were unable to resume work due to problems caused by cancer. However, more research is needed to identify the impact of AML on productivity losses. The exact impact of AML and its treatments on productivity is currently unknown. In order to assess the broader societal impact of AML, it is important that productivity losses are explicitly measured in addition to quality of life

ESSENTIAL ELEMENTS IN COST-EFFECTIVENESS ANALYSIS OF PERSONALIZED MEDICINE STRATEGIES

Although this thesis performed cost-effectiveness analyses of personalized medicine strategies for acute myeloid leukemia, it also identified generic methods applicable to other diseases. Since the use of personalized medicine is increasing in many other diseases (other cancer types, cardiovascular disease and mental health), there is a need for generalized methods of analysis. In general, four steps can be identified in the cost-effectiveness analyses of personalized medicine strategies:

1. Exploration of how treatment will change due to the application of personalized medicine strategies;
2. Assessment of the impact of the treatment changes on health outcomes and costs;
3. Assessment of the performance of the diagnostic test;
4. Assessment of the impact of the test strategy on health outcomes and costs.

These four steps are combined in the formula developed in chapter 8. The formula enables the assessment of the cost-effectiveness of the combination of the test and treatment strategy, but also report the separate effects of the test and treatment to improve the interpretation of the results. All four steps are interconnected and should therefore not be conducted independently. The following paragraphs describe some crucial components of each of the four steps and identifies areas for future research to improve the methodology of cost-effectiveness analyses of personalized medicine strategies.

Exploration of how treatment changes due to the application of personalized medicine strategies

The aim of personalized medicine strategies is to select patients who benefit from a specific treatment. New insights about the prognostic impact of, for example, a specific molecular abnormality may result in changes in the algorithm for selecting patients and the subsequent treatment. However, treatment selection is often based on several other factors, such as patient characteristics, clinical factors and other molecular abnormalities, as well. Cost-effectiveness analyses of personalized medicine strategies should therefore not assume that treatment changes for all patients with the specific abnormality. This finding can be clearly explained by the case study in AML (Chapter 8). The personalized medicine strategies for AML aimed to select patients who benefit from allogeneic hematopoietic stem cell transplantations. These transplantation were restricted to patients with an intermediate or high risk of relapse. However, not all patients were eligible for these transplantations due to comorbidities or a low performance status of the patient or the absence of a suitable donor. Consequently, treatment only changed for a subset of the newly identified patients.

A correct assessment of the proportion of patients for whom treatment changes due to adjustments in the personalized medicine strategy is crucial for the estimation of the cost-effectiveness of the strategy. It was shown that the ICER increases substantially if treatment changes for a smaller group of patients. Consequently, the cost-effectiveness of personalized medicine strategies will probably be overestimated if analyses assume that treatment changes for all patients in the newly identified subgroup.

The proportion of patients with a treatment change also depends upon the uptake of the personalized medicine strategy in clinical practice. A limited uptake may be related to the application of the test or the choice of treatment according to the test results. Both examples have been found regarding the use of erlotinib for lung cancer patients with *EGFR* mutations (279). Despite convincing evidence about the benefits of erlotinib in this subgroup of patients, not all hospitals in the Netherlands tested for *EGFR* mutations in lung cancer patients. Furthermore, even if the test was performed, treatment choice was not always in accordance with the test results. The limited uptake of the treatment guideline may reduce the proportion of patients with a treatment change and increase the ICER. The limited application of the test will only impact the proportion with a treatment change if the reasons for deciding to use the test are not random. With respect to cost-effectiveness analyses of personalized medicine strategies, it is recommended to assess the full potential costs and benefit of the personalized medicine strategy by assuming perfect implementation of both the test and treatment approach. Furthermore, scenario analyses are needed to assess the real-world cost-effectiveness of the strategies while assuming imperfect implementation of both the test and treatment strategy.

Finally, the proportion of patients whose treatment changes also depends upon the qualitative performance of the diagnostic test. Assuming that a positive test results in a treatment change, the proportion decreases for a lower sensitivity and increases for a lower specificity. The contrary holds once a negative test results in a treatment change.

The assessment of the impact of treatment changes on health outcomes and costs

The developed formula presented in chapter 8 shows that only health outcomes and costs of patients with a treatment change need to be included in the analysis. In theory, time and money can be saved by restricting the data collection and analysis of health outcomes and costs to this selected subgroup of patients. However, it is often not feasible to identify the patient group with a treatment change prior to the start of the study. Furthermore, in the continuous developing field of personalized medicine, it might be more efficient in the long run to collect data of all patients. New evidence about prog-

nostic subgroup may adjust the personalized medicine strategy and require additional information about health outcomes and costs.

Existing methods of cost-effectiveness analyses, including piggy-back studies and decision-analytic modeling, can in principle be used to assess the impact of personalized medicine strategies on health outcomes and costs. However, some decisions regarding type of data sources and decision-analytic model may differ from standard cost-effectiveness analyses of new drugs. The most commonly used decision-analytic model in cost-effectiveness analyses, the Markov model, can be used for the assessment of personalized medicine strategies, but it is less flexible to incorporate future adjustments to personalized medicine strategies than a discrete-event simulation. Therefore, the discrete-event simulation was used in this thesis to assess the cost-effectiveness of personalized medicine strategies in AML. Nevertheless, the discrete-event simulation has also some disadvantages, including the computational and analytical complexity and the dependency on individual patient level data. Model selection for future cost-effectiveness analysis of personalized medicine strategy should result from a deliberated choice to balance complexity, flexibility and feasibility.

The feasibility of a specific model depends largely on the availability of the data. The involvement of health economist at the design phase of new clinical studies may improve access to adequate information by including relevant health economic outcomes such as costs and quality of life. Notwithstanding, the actual data collection is often restricted due to strict budgets. It is expected that clinical studies of personalized medicine strategies are more extensive than studies for new drug entities, because the assessment of both the test and treatment is not feasible in standard clinical trial designs. Therefore, new types of study design are currently being used to assess the impact of personalized medicine strategies. Four of these new randomized controlled trial designs are, respectively, the biomarker-stratified design, the enrichment design, the biomarker-strategy design and the adaptive trial design (280-282). The biomarker-stratified design stratifies patients according to test result and subsequently randomizes both patients with a positive and negative test result to the different treatment arms. The enrichment design initially includes all patients. However, only patients with a positive test outcome will be randomized to the different treatments. In the biomarker-strategy design, patients are randomized between a uniform and stratified treatment according to the test results (280). The adaptive trial design is an extended version of the biomarker-strategy design and divides the trial period into two phases. The first phase is aimed at selecting the most appropriate test to stratify treatment. Nevertheless, in that phase, treatment is already randomized between uniform and stratified treatment using a gold standard to stratify. In the second phase, stratification is based upon the most appropriate test

according to the results of phase 1 (282). All described trial designs have advantages and disadvantages and there is currently no general agreement yet regarding the optimal trial design. At this moment, the choice of trial design largely depends upon the aim of the study and the balance between the required evidence and the feasibility of collecting this evidence in a timely manner (280,281).

Regardless of the challenge of the choice of an appropriate design of clinical trials for personalized medicine strategies, it can be challenging to recruit a sufficiently large sample. Although the total required sample size of the trials might be smaller due to a larger expected efficacy, the total source population of eligible patients is also much smaller. Accordingly, the recruitment of patients may take much longer, which will increase the development time of the new medicines and technologies. In order to solve this problem, international collaborations between research centers worldwide are needed to speed up the inclusion of the required patient numbers. Notwithstanding, worldwide collaboration might be practically difficult if the standard clinical practice differs among countries. Another disadvantage of worldwide collaboration is that only one treatment modality may be studied at a time. This problems is solved by new designs like the 'pick-a-winner' design which start with the inclusion of several treatment modalities and use interim analyses to select the treatment with the most promising results for further evaluation (283). However, a major disadvantage of this design is the high chance for incorrect decision about treatment (dis)continuations.

All new designs are mainly developed to show the efficacy of personalized medicine strategies. However, randomized clinical trials have strict in- and exclusion criteria to limit the bias of other factors such as comorbidities and poor performance status in the assessment of the treatment effect. The disadvantage of the strict in- and exclusion criteria is the lack of generalizability towards the real clinical practice. Patients excluded from clinical trials are often older, less fit and have more comorbidities than included patients (284). An assessment of the treatment effect in daily practice, the effectiveness, requires additional data. This data might be derived from patient registries which are increasingly being set up to assess the cost-effectiveness of new medicines in daily practice. However, it has been shown that the selection of an adequate control group is difficult due to the absence of randomization in daily practice. Treatment choice is largely driven by patient and disease characteristics (285). Consequently, cost-effectiveness analyses using real word data are often biased due to differences in patient and disease characteristics between treatment groups. A historical cohort might be a better comparator as a group with similar patient and disease characteristics can be selected. However, not infrequently, historical controls are not available for the assessment of personalized medicine strategies because tests for biomarkers have not been performed in the past.

Patient registries should address this problem to enable real-world cost-effectiveness analyses of personalized medicine strategies. A possible solution is to expand existing registries with the storage of patient material for future analyses if feasible regarding legal and ethical boundaries.

The assessment of the performance of diagnostic tests

The test strategy is another critical component in the evaluation of personalized medicine strategies in addition to the consequences of a treatment change. Since the qualitative performance of the test largely influences the proportion of patients with a treatment change, it is important that the performance of the test is adequately assessed in cost-effectiveness analysis of personalized medicine strategies. Within this perspective, it is important to recognize that different definitions of performance can be used. The strict definition of performance assesses whether the test correctly identifies a (molecular) abnormality (286). The disadvantage of this definition is that it assumes that a current test is used as a gold standard, while it is questionable whether a gold standard can be defined for newly developed technologies. A broader definition of performance relates to the performance of risk prediction, for example whether tests correctly distinguish responders from non-responders before the start of the treatment (287). This broader definition is better applicable to personalized medicine strategies because treatment choice depends on several factors including age, performance status and disease characteristics. All these different measures can be included in the prediction algorithm and adjustments to the algorithm can also easily be assessed.

The assessment of the impact of the test strategy on health outcomes and costs

The impact of the test strategy on the cost-effectiveness of personalized medicine strategies is determined by the costs and the application of the test. The most obvious effect is that the cost of a personalized medicine strategy increase with higher costs of the test. The application of the test refers to the decision about the patient group to be tested, i.e. is the test applied to all patients with a specific disease or only a subset of these patients? This decision determines the size of the subgroup for whom treatment changes as proportion of the total tested population. This thesis showed that the ICER increases for smaller subgroups. Since both the costs and the application of the test can be influenced by the manufacturer, scenario analyses regarding different costs and applications are very informative to select a cost-effective testing strategy. These adjustments to the testing strategy may not only be beneficial for the manufacturer, but also for patients as it is more likely that personalized medicine strategies will be reimbursed if these are cost-effective.

In this thesis, it was assumed that testing only influences health outcomes by means of a treatment change. Consequently, differences in health outcomes were not observed for patients without a treatment change. However, there may be situations in which the test itself has an impact on health outcomes. For example, invasive tests like colonoscopy may reduce short-term quality of life due to discomfort and pain. Other tests may have more severe complications such as infection or death. Furthermore, tests may yield benefits or harms for patients if uncertainty about the diagnosis and prognosis is resolved regardless of any changes in treatment (288). Future studies should assess how these additional benefits and harms of personalized medicine strategies can be included in cost-effectiveness analyses and reimbursement decisions. A critical element for these studies is whether all benefits and harms can be adequately measured by the QALY concept. Short-term changes in quality of life have little impact on the QALY, but are important if patient preferences are also part of reimbursement decisions. Furthermore, it is unknown whether the value of knowing (resolving uncertainty about diagnosis and prognosis) is already included in the QALY concept or whether additional outcome measures are needed.

A complicating factor in the assessment of the personalized medicine strategies is that it is expected that new tests, like whole genome sequencing of tumor samples and full body imaging scans, are able to detect several abnormalities at a time. The result of the test is no longer dichotomous, i.e. presence or absence of one abnormality, but consists of a combination of abnormalities. Cost-effectiveness analyses of these new tests require the assessment of different treatment options for all possible combinations of abnormalities. Besides an expansion of the number of strategies to be evaluated, whole genome sequencing of tumor and full body imaging scans have a substantial risk of incidental findings due to the broad application of the tests (289). Incidental findings are results which were not anticipated when ordering the test. These findings require additional testing and treatment to adequately treat the abnormality detected (290). It should be discussed whether and how the costs and effects of these incidental findings can be incorporated in cost-effectiveness analyses of personalized medicine strategies.

OTHER RELEVANT CONSIDERATIONS FOR SUCCESSFUL IMPLEMENTATION OF PERSONALIZED MEDICINE STRATEGIES.

It is expected that a more standardized method to analyze and report the cost-effectiveness of personalized medicine strategies improves the implementation of the strategies as the consequences of the strategies are more transparent. However, better methodological guidelines are not sufficient for a successful implementation of new

personalized medicine strategies. Adaptation in the development and reimbursement of the personalized medicine strategies are also required.

More valuable personalized medicine strategies might be introduced to the market if the development of these strategies is guided by early cost-effectiveness analyses (41,291). These early analyses can identify the potential of the strategy by evaluating the required targets in terms of costs of the test and treatment, application of the test and treatment effectiveness. The strategy can be adjusted early in the development process without wasting resources on the development of unsuccessful strategies. Furthermore, the data collection can be improved as the early assessment may also identify the most influential parameters for which additional data is required.

The reimbursement of personalized medicine strategies should be based upon integral financing of both the test and treatment. Although the diagnosis-related groups aim for a more integrated financing system by combining diagnosis and treatment in one declaration form, it is not adequately working in current practice. Each hospital division is still responsible for their own budget and personalized medicine strategies often lead to additional costs in one division, while the benefits are accrued by another division. For example, laboratories need to invest in new genetic tests, while the clinical department saves money by applying a less costly treatment. A more integral financing system might improve the incentives for laboratories to implement new technologies.

Besides an integral method of financing personalized medicine strategies, it is recommended to apply value-based pricing according to the added value of both the test and targeted treatment. The added value of the targeted treatment can be determined by the (cost-)effectiveness of the new treatment in the selected subgroup, while tests have added value if new subgroups are identified or if the identification of existing subgroups is improved. This type of value-based pricing allow higher reimbursement levels for treatments targeted to newly identified subgroups to cover the costs of testing as well. A standardized reimbursement level for all personalized medicine strategies is either not sufficient for strategies related to newly identified subgroups or overcompensates strategies for subgroups already identified in clinical practice. Consequently, value-based pricing stimulates innovation for both existing and new subgroups while allowing a more sustainable health care.

Chapter 10

Summary

Samenvatting

Dankwoord

Phd portfolio

About the author

References

SUMMARY

Treatment for cancer is nowadays moving towards a more personalized approach, because it has been shown that only a selective group of patients responds to the available treatments. Furthermore, disease susceptibility and prognosis also differ from patient to patient. The selection of treatment according to individual patient characteristics is called 'personalized medicine' and may be based upon both genetic and non-genetic information. It is expected that the number of personalized medicine strategies will largely increase in the future due to the growing evidence about the association between genetic aberrations and prognosis or treatment response. Since techniques to identify genetic aberration can be very expensive, cost-effectiveness analyses of personalized medicine strategies are required to allow for a sustainable health care in the future. This thesis assess the cost-effectiveness of personalized medicine strategies in acute myeloid leukemia (AML) and aims to contribute in methodology for strategies in other areas.

AML is a specific type of leukemia which is characterized by a proliferation of immature myeloid cells (blasts) in the bone marrow. The proliferation of blasts reduces the development of normal blood cells which lead to increased risk of bleedings and infections and make patients feel tired and weak. The rapid increase of blasts in patients with AML requires immediate treatment to reduce the number of blasts and restore the normal blood functioning. Intensive treatments, including high-dose chemotherapy and stem cell transplantations, are required to cure the disease. Unfortunately, these treatments are often only feasible in younger patients, resulting in a better prognosis for these patients. The 5-year overall survival rate is 55% and 9% in patients aged 18-44 and 65-74 years, respectively. Another important prognostic factor besides age is the presence of cytogenetic and molecular aberrations. Many different aberrations have been identified and some of these are already defined as a distinct disease entity by the World Health Organization.

There is high potential for personalized medicine strategies in AML due the to the heterogeneity nature of the disease. At this moment, a targeted treatment is only available for one AML subgroup. The application of this targeted treatment is already standard clinical practice and dramatically improved the prognosis of that subgroup. The current options to further personalize AML treatment are only available for younger aged patients (age < 60 years) and include the selection of treatment according to expected risk of relapse. Patients with a low risk of relapse can be cured with a less intensive treatment (high-dose chemotherapy instead of (allogeneic) hematopoietic stem cell transplantation) than patients with a high risk of relapse. Information about the prognostic impact of newly identified subgroups in AML resulted in a reclassification of patients from the

intermediate risk group to either the low or high risk group. This thesis assessed the impact of these reclassifications on health outcomes and costs.

Estimation of input parameters

The analyses started by measuring the costs of all treatments in acute myeloid leukemia (Chapter 2). The costs were distinguished in three treatment phases: induction treatment, post-remission treatment and 1-year follow-up. Post-remission treatment consisted of high-dose chemotherapy, autologous hematopoietic stem cell transplantation (HSCT) and allogeneic HSCT from a sibling or matched unrelated donor (MUD). Costs were based on the resource use of all adult patients diagnosed with de novo primary or secondary AML who started with induction chemotherapy in 2008 or 2009 in three university hospitals in the Netherlands. Treatment for AML is very expensive; the costs of induction chemotherapy are about €45,000 and the costs of post-remission treatment ranges between €33,277 and €82,041. A large proportion of these costs is caused by the long hospital stay (about four weeks) during treatment. An allogeneic HSCT from a MUD is the most expensive treatment option due to the costs of donor searching, HLA-typing and additional diagnostic tests. The treatment of graft-versus-host disease as a complication of an allogeneic HSCT resulted in significantly higher follow-up costs after these procedures.

The most common effect measure in cost-effectiveness analyses is the quality-adjusted life year (QALY), a combined measure of mortality and morbidity. This measures weights the gain in life years by a utility value for the quality of life. Perfect health is represented by a utility value of 1 and death with a utility value of 0. These utility values are often derived from generic preference-based questionnaires such as the EQ-5D. These questionnaires are applicable to all diseases and the utility values were calculated from preferences of the general public. Many clinical studies, however, use disease-specific instruments which do not incorporate these preferences in the scoring algorithms. Consequently, the results of these studies cannot be used to estimate quality of life utilities. This problem was also observed in acute myeloid leukemia. Therefore, two different methods were applied in this thesis to derive utilities from a cancer-specific quality of life questionnaire (QLQ-C30). These methods were also applied to disease-specific questionnaire in two other disease areas, multiple sclerosis and rheumatoid arthritis, to find generic pattern for these methodology.

Chapter 3 describes the application of the first method to estimate utility values from disease-specific quality of life questionnaire, which is called mapping. Mapping predicts the EQ-5D score based upon responses to the questionnaire. The prediction is derived from regression analysis and is therefore only feasible if the EQ-5D and the disease-

specific questionnaire are included in the same study. With respect to the QLQ-C30, the mapping algorithm was developed in a sample of patients with multiple myeloma and tested in sample of patients with Non-Hodgkin lymphoma. Model selection was based upon statistical significance of predictors, logical consistency, parsimony and predictive ability. The selected mapping algorithm depends on the overlap of the items on the disease-specific instrument with the EQ-5D and the population in which the algorithm was developed. Ideally mapping functions are developed in a representative sample of all patients to improve external validity. A comparison between the mapping algorithms for the three disease-specific questionnaires showed the best performance for the QLQ-C30. This finding was in line with the expectations, because the QLQ-C30 had the largest correlation with the EQ-5D. The mapping algorithm for the QLQ-C30 as developed in the Multiple Myeloma patient sample was able to predict utilities in the Non-Hodgkin's sample. Furthermore, the predicted utilities for the four performance states followed the same pattern as the observed EQ-5D utilities. However, the differences in utility scores between the performance states were slightly smaller.

The second method to calculate utility values for disease-specific questionnaires is applied in chapter 4 and include the direct valuation of disease-specific health states. The direct valuation was performed for the same disease-specific questionnaires as for which mapping algorithm were developed. The aim of the study was to assess the comparability of utility values from these disease-specific questionnaires to utility values from generic questionnaires. The disease-specific questionnaires were first reduced in content, because respondents can maximal value nine items. Reduction was based on three criteria: i) fit to the Rasch model, ii) standard psychometric criteria and iii) clinical relevance. Once the questionnaires were reduced, a subset of all possible health states was selected for valuation. Respondents from the general public valued these health states by means of the time-trade-off method similar to the protocol used in the valuation of the EQ-5D. The preference values observed for the selected health states were used to estimate values for all potential health states through statistical modeling. The results of the study showed significant higher utility values for all three disease-specific questionnaires compared to utility values derived from generic preference-based questionnaires. The higher average values are caused by the smaller range of utility values indicating that the disease-specific questionnaires do not capture very poor health. The preference-based instrument for cancer (QLQ-PBM) appeared to be more sensitive to small changes in health as a substantial proportion of patients reported problems on this instrument while reporting no problems on the EQ-5D. Furthermore, both the QLQ-PBM and EQ-5D were able to detect differences between patients with and without side effects. No evidence was found for an improved discriminative ability or responsiveness of the QLQ-PBM compared to the EQ-5D. In fact, both instruments were able to detect

differences over time, but the absolute difference in utility was consistently higher for the EQ-5D.

A comparison of the two methods to estimate utility values for the QLQ-C30 showed that utility values resulted from the mapping algorithm were best comparable to those from the EQ-5D. However, the algorithm excluded items which were not adequately captured by the EQ-5D. Therefore, the use of the QLQ-PBM is preferred if these excluded scales are considered essential for measuring quality of life in patients with acute leukemia. Otherwise, the EQ-5D, or the mapping algorithm if the EQ-5D is absent, should be used in cost-effectiveness analyses to allow comparisons with other diseases. An assessment of the validity of both the EQ-5D and the QLQ-PBM In acute leukemia patients was required to judge which questionnaire can best be used in the cost-effectiveness analyses of personalized medicine strategies in AML. This assessment was feasible with the results of a quality of life study in acute leukemia survivors in which both questionnaires were included. The results of this study are described in chapter 6. A newer version of the EQ-5D, with five answer levels for each domain instead of three, was included as it was assumed that this version was more sensitive to small changes in health. According to this study, no problems were indicated regarding the feasibility, validity and reliability of the EQ-5D in acute leukemia patients. Both questionnaires had a good internal consistency and only six and seven out of the 111 respondents did not fully complete the EQ-5D and QLQ-PBM, respectively. Furthermore, both instruments were able to distinguish between severity levels. However, a substantial larger proportion of patients reported full health on the EQ-5D compared to the QLQ-PBM.

The quality of life study was initially set up to assess the impact of AML and its treatment on health-related quality of life (HRQOL) by comparing the HRQOL of AML survivors with the HRQOL in the general population as described in chapter 5. Questionnaires were sent to patients diagnosed with acute leukemia between 1999 and 2011 at a single academic hospital and still alive in 2012. It was found that AML survivors had a worse quality of life than the general population. Problems were more frequently reported regarding all functioning scales, pain, dyspnea, fatigue, appetite loss and financial difficulties. The impaired quality of life was especially found in patients without a paid job. This is an important problem as the group accounts for about 25% of all AML survivors in our study. The study also indicates factors possibly associated with a poor HRQOL. These factors include lack of social support, allogeneic HSCT as post-remission treatment and younger age. However, due to the insufficient power of the study no definite conclusions can be drawn.

Cost-effectiveness analyses

A decision-analytic model was developed and validated to assess the cost-effectiveness of personalized medicine strategies in acute myeloid leukemia. The development and validation was an iterative process and the results were continuously checked with clinical experts and input data for face validity and internal validity. Detailed information about the development and validation can be found in chapter 7. A discrete-event simulation was developed to adequately capture the heterogeneity of the disease and the impact of several disease and patient characteristics on important health outcomes such as complete remission rate, risk of relapse and death. A fictive dataset was simulated with individual disease paths. For each patient, the disease paths was determined by the estimation of the time to concurrent events from survival functions including relevant patient and disease characteristics. The event with the shortest time to occurrence was always selected for each patient. Both patient and disease characteristics as well as relevant clinical outcome measures of the final developed model were compared with the original data. No striking differences were found between the model and the original data indicating a good internal validity of the model. The model results were also in line with the reported survival in other clinical trials. However, the generalizability of the model to a broader patient population has not been proven yet.

As described before, the personalized medicine strategies under study consist of reclassification of patients to either the low or high risk group. Separate cost-effectiveness analyses were performed for the two reclassification types and reported in chapter 8. As a consequences of the reclassification, treatment changes for the new risk groups; a less intensive treatment is administered to patients with a low risk and a more intensive treatment to patients with a high risk. The cost-effectiveness analysis was first restricted to the assessment of the consequences of the treatment changes while neglecting the costs of identifying the new subgroups, because many published studies used this restricted approach. It was found that the less intensive treatment for low risk patients resulted in better health outcomes at lower costs. The new treatment was therefore dominant for this subgroup. Health was also improved by the more intensive treatment for high risk patients, but at substantial higher costs. The incremental cost-effectiveness ratio (ICER) of the treatment change for that subgroup was €130,065 per QALY gained. The inclusion of the consequences of the test resulted in higher ICERs for both personalized medicine strategies. It was found that the severity of the underestimation of the ICER depends upon the costs of the test, the incremental effects of the treatment change and the prevalence of the subgroup with a treatment change. According to these findings, a formula was developed that enables the separate assessment of the consequences of the test and treatment approach in personalized medicine strategies.

Recommendations for future research

Chapter 9 discusses the main findings and implications of this thesis. It was shown that it is cost-effective to treat patients with a low risk of relapse with a less intensive treatment. However, more cost-effective treatments are required for patients with a high risk of relapse. The new treatment for these patients is very costly and has negative consequences in terms of quality of life. Besides improved treatment for this specific subgroup, targeted treatments for other AML subgroups are also required. Other improvements in the treatment of AML may be achieved by better supportive care programs which may improve quality of life. Furthermore, costs can largely be reduced if patients can be treated in an outpatient setting. More effective, but less toxic, treatment might enable this outpatient setting.

Preferably, new clinical studies include quality of life as a secondary outcome. Although it seems that the EQ-5D is valid for acute leukemia patients, it is recommended to include both the disease-specific QLQ-C30 and generic EQ-5D to test the validity of the instruments in larger longitudinal studies. Furthermore, both direct and indirect costs should be measured in future studies to adequately assess the impact of new treatment strategies on the total costs.

The findings from the case-study in AML also contribute to the methodology of cost-effectiveness analyses of personalized medicine strategies in other diseases areas. It is recommended to always include the costs of testing in cost-effectiveness of personalized medicine strategies to avoid biased results. The use of the formula may enable an efficient approach to add the costs of testing to the analysis. Data collection and analysis of health outcomes and costs can still be restricted to the subgroup with a treatment change. Standardized use of the formula also improves the interpretation of the results by better comparability between studies. Furthermore, the formula can be used early in the development process to guide further development and data collection.

SAMENVATTING

De behandeling van kanker wordt in toenemende mate afgestemd op de individuele kenmerken van de patiënt. De achterliggende reden hiervoor is dat slechts een beperkte groep patiënten reageert op de beschikbare behandelingen en die groep patiënten is steeds beter op voorhand te identificeren. Naast response kunnen ziekteprognose en de kans op het ontwikkelen van ziektes ook sterk verschillen tussen individuen. Verschillende soorten informatie kunnen worden gebruikt om de juiste behandeling te selecteren, zoals leeftijd, comorbiditeiten, afwijkende laboratoriumwaarden en moleculaire afwijkingen. Aangezien steeds meer verbanden worden gevonden tussen moleculaire afwijkingen en ziekteprognose of behandeleffect, wordt verwacht dat het aantal geïndividualiseerde behandelstrategieën sterk zal toenemen in de toekomst. Het onderzoeken van de kosteneffectiviteit van deze nieuwe strategiëen is daarom essentieel om een betaalbare gezondheidszorg te blijven garanderen in de toekomst. Dit proefschrift onderzoekt de kosteneffectiviteit van verschillende geïndividualiseerde behandelstrategieën in acute myeloïde leukemie (AML). Naast belangrijke bevindingen voor AML levert het proefschrift ook aanbevelingen voor de methodologie van kosteneffectiviteitsanalyses van geïndividualiseerde behandelstrategieën in andere ziektegebieden.

AML is een specifieke vorm van leukemie dat wordt gekenmerkt door een ongecontroleerde toename van onvolgroeide myeloïde cellen (blasten) in het beenmerg. De sterke toename van de blasten beperkt de ontwikkeling van normale bloedcellen waardoor patiënten een verhoogd risico hebben op bloedingen en infecties. Aangezien AML een zeer heterogene ziekte is met veel verschillende cytogenetische en moleculaire afwijkingen is er veel potentie voor het toepassen van geïndividualiseerde behandelstrategieën. Op dit moment is slechts één gerichte behandeling beschikbaar voor een specifieke AML subgroup. Deze behandeling heeft de prognose voor de die subgroup aanzienlijk verbeterd. Met de huidige kennis kan de behandeling van AML alleen verder geïndividualiseerd worden door de behandeling af te stemmen op de verwachte kans op terugkeer van de ziekte (relapse). Patiënten met een kleine kans op relapse kunnen worden behandeld met een minder intensieve behandeling (hoge dosis chemotherapie in plaats van een (allogene) stamceltransplantatie) dan patiënten met een hoge kans op relapse. Informatie over de prognostische impact van nieuwe geïndenticeerde subgroepen van AML heeft geresulteerd in een herclassificatie van patiënten in de verschillende risicogroepen. De gevolgen van deze herclassificatie op gezondheidsuitkomsten en kosten zijn onderzocht in dit proefschrift.

Achterhalen van de inputs voor de kosteneffectiviteitsanalyse

De eerste stap in de analyse was het bepalen van de kosten van de verschillende behandelingen voor AML (hoofdstuk 2). Alle kosten werden onderscheiden in drie behandelfases: inductie behandeling, post-remissie behandeling en 1-jaar follow-up. De post-remissie behandeling bestond uit hoge-dosis chemotherapie, autologe stamceltransplantie en een allogene stamceltransplantatie van een sibling (broer/zus) of een donor afkomstig uit een donor bank. Alle kosten waren gebaseerd op het zorggebruik van volwassen patiënten die in 2008 of 2009 waren gestart met een inductie behandeling in drie universitaire ziekenhuizen in Nederland. Uit de studie bleek dat de behandeling voor AML erg duur is. De kosten van één inductie chemokuur bedroegen ongeveer €45.000 en de kosten voor de post-remissie behandeling variëren tussen de €33.277 en €82.041. Een aanzienlijk deel van de kosten wordt veroorzaakt door de langdurige ziekenhuisopname (ongeveer vier weken) tijdens de behandeling. Een allogene stamceltransplantatie van een ongerelateerde donor (afkomstig uit de donorbank) is de duurste behandeling. De hogere kosten worden veroorzaakt door de kosten voor het zoeken van een donor, HLA-typering en aanvullende diagnostische testen. De kosten voor de follow-up na allogene stamceltransplantaties waren significant hoger dan andere follow-up kosten door de behandeling van graft-versus-host-disease als complicatie van de stamceltransplantaties.

De meest gebruikte effectmaat in kosteneffectiviteitsanalyes is het voor kwaliteit gecorrigeerde levensjaar (quality-adjusted life year oftewel QALY). De QALY combineert mortaliteit en morbiditeit in één maat door elk gewonnen levensjaar te wegen met een utiliteit voor de kwaliteit van dat leven. Perfecte gezondheid heeft een utiliteit van 1 en dood een utiliteit van 0. Deze utiliteiten zijn vaak afgeleid van generieke vragenlijsten zoals de EQ-5D om vergelijkingen tussen ziektes te kunnen maken. In veel klinische studies, ook in AML, wordt deze vragenlijst echter niet altijd meegenomen, maar wordt gebruik gemaakt van een ziektespecifieke vragenlijst. Oorspronkelijk was het niet mogelijk op basis van deze ziektespecifieke vragenlijsten kwaliteit van leven utiliteiten. Het gevolg hiervan is dat de QALY niet gebruikt kan worden als uitkomstmaat in kosteneffectiviteitsanalyses. Om dit probleem op te lossen, zijn in dit proefschrift twee verschillende methoden toegepast om utiliteiten te achterhalen voor een kankerspecifieke vragenlijst (QLQ-C30). Deze methoden zijn ook toegepast op ziektespecifieke vragenlijsten in twee andere ziektegebieden (multipele sclerose en rheumatoïde arthritis) om generieke uitspraken te kunnen doen over de gebruikte methodologie.

Hoofdstuk 3 beschrijft de toepassing van de eerste methode, 'mapping' genoemd. Bij mapping wordt de score op de EQ-5D voorspeld op basis van de items van de ziektespecifieke vragenlijst. Deze voorspelling is alleen mogelijk als patiënten zowel de EQ-5D

als de ziektespecifieke vragenlijst hebben ingevuld. Voor de QLQ-C30 is het mapping algoritme afgeleid uit een studie met patiënten met multipel myeloom en getest in patiënten met non-Hodgkin lymfoom. Verschillende criteria, zoals statische significantie, logische consistentie, eenvoud en voorspellend vermogen, zijn gebruikt om de voorspellend algoritme te ontwikkelen. Ult de studie bleek dat de kwaliteit van het algoritme afhankelijk is van de populatie die gebruikt is voor de voorspelling en de overlap tussen de items op de ziektespecifieke vragenlijst en de EQ-5D.

De tweede methode voor het bepalen van utiliteiten voor ziektespecifieke vragenlijsten is toegepast in hoofdstuk 4 en betreft een directe waardering van ziektespecifieke gezondheidstoestanden. De eerste stap in de analyse was het inkorten van de ziektespecifieke vragenlijsten, omdat respondenten maximaal negen verschillende items kunnen waarderen. De volgende criteria werden gebruikt voor de selectie van de items: i) overeenstemming met het Rasch model, ii) standard psychometrische criteria en iii) klinische relevantie. Respondenten uit het algemeen publiek hebben een subset van gezondheidstoestanden gewaardeerd met de time trade-off methode die ook gebruikt was in de waardering van de EQ-5D. De utiliteiten van niet-gewaardeerde gezondheidstoestanden zijn vervolgens geschat met behulp van statistische modelleertechnieken. De utiliteiten voor de ziektespecifieke vragenlijsten bleken significant hoger te zijn dan die van de EQ-5D. Dit kan mogelijk verklaard worden door de kleinere range van utiliteiten bij ziektespecifieke vragenlijsten. Het lijkt dat zeer slechte gezondheid, als gevolg van bijvoorbeeld comorbiditeit, niet wordt opgepikt door de ziektespecifieke vragenlijsten. Een vergelijking tussen de utiliteiten van de EQ-5D en de QLQ-PBM laat zien dat kleine verschillen in gezondheid beter worden geïdentificeerd door de QLQ-PBM. Beide instrumenten waren echter in staat om een onderscheid te maken in de kwaliteit van leven bij patiënten met en zonder bijwerkingen. Er was ook geen bewijs gevonden voor een verbeterde responsiviteit van de QLQ-PBM.

Een vergelijking van de twee methoden om utiliteiten te berekenen voor de QLQ-C30 laat zien dat mapping leidt tot utiliteiten die het beste vergelijkbaar zijn met die van de EQ-5D. Het nadeel van mapping is echter dat items die niet adequaat worden opgepikt door de EQ-5D ook niet zullen worden opgenomen in het mapping algoritme. Deze items kunnen wel zijn opgenomen in de QLQ-PBM. Op basis van deze informatie is de vraag welk instrument het best gebruikt kan worden voor het meten van kwaliteit van leven utiliteiten bij patiënten met acute leukemia. In principe wordt de EQ-5D aangeraden voor kosteneffectiviteitsanalyses omdat de resultaten vergeleken kunnen worden met andere ziektes. Een noodzakelijke voorwaarde is dan wel dat de EQ-5D valide is voor de desbetreffende patiëntenpopulatie. De validiteit van zowel de EQ-5D als de QLQ-PBM bij acute leukemie is onderzocht in hoofdstuk 6. Zowel de EQ-5D als de QLQ-PBM

hadden een goede interne consistentie en slechts zes en zeven respondenten hadden respectievelijk de EQ-5D of de QLQ-PBM niet volledig ingevuld. Daarnaast waren beide vragenlijsten in staat om een onderscheid te maken naar ernst van de ziekte. Wel werd perfecte gezondheid veel vaker gerapporteerd op de EQ-5D dan de QLQ-PBM. Op basis van deze studie is geen reden gevonden om te twijfelen aan de validiteit van de EQ-5D.

De cross-sectionele kwaliteit van levenstudie was in eerste instantie opgezet om de kwaliteit van leven te meten bij patiënten die in het verleden (tussen 1999 en 2011) waren gediagnosticeerd met AML. Door deze resultaten te vergelijken met de kwaliteit van leven in de algemene populatie kon inzicht worden gekregen in de invloed van AML en de bijbehorende behandelingen op de kwaliteit van leven. De AML patiënten rapporteerden een slechtere kwaliteit van leven dan de algemene populatie. Ze hadden vaker problemen met verschillende vormen van functioneren en ervaarden vaker pijn, dyspneu, vermoeidheid, gebrek aan eetlust en financiële problemen. Een lagere kwaliteit van leven was voornamelijk gevonden bij jongere patiënten zonder betaalde baan. Dit is een belangrijk probleem want in totaal was 25% van de respondenten niet in staat om te werken als gevolg van AML. Uit de studie bleek ook dat een allogene stamceltransplantatie, gebrek aan sociale steun en een lagere leeftijd geassocieerd waren met een slechtere kwaliteit van leven. Vanwege het relatief kleine aantal respondenten kunnen hier echter nog geen definitieve conclusies aan verbonden worden.

Kosteneffectiviteitsanalyses

Voor het berekenen van de kosteneffectiviteit van geïndividualiseerde behandelstrategieën in AML is een analytisch beslismodel ontwikkeld en gevalideerd. De ontwikkeling en validatie was cyclisch proces, waarbij de resultaten continu werden gechecked met klinische experts en input data. Gedetailleerde informatie over dit proces is te vinden in hoofdstuk 7. Het model was een discrete-event simulatie om de heterogeniteit van AML goed mee te kunnen nemen in de analyses. Een fictieve dataset was gesimuleerd op basis van individuele ziektepaden. Deze individuele paden volgden uit schattingen van de tijd tot verschillende events waarbij rekening was gehouden met patient- en ziektekenmerken. Voor iedere patient werd steeds de tijd tot het kortste event geselecteerd. Zowel de patient- en ziektekenmerken als de relevant klinische uitkomstmaten in de gesimuleerde dataset kwamen overeen met de originele data. Daarnaast was de overleving uit het model ook vergelijkbaar met de gerapporteerde overleving in klinische trials die niet gebruikt waren als input voor het model. De generaliseerbaarheid van het model naar een bredere patiëntenpopulatie was echter nog niet bewezen. De geindividualiseerde behandelstrategieën die geëvalueerd zijn in dit proefschrift betroffen een herclassificatie van patiënten uit de gemiddelde risico groep naar zowel de lage als de hoge risicogroep. Aparte kosteneffectiviteitsanalyses zijn uitgevoerd voor deze twee

herclassificatie opties (zie hoofdstuk 8). Een belangrijk gevolg van de herclassificatie was een verandering in de behandeling: een minder intensieve behandeling werd gegeven aan patiënten met een laag risico en een intensievere behandeling aan patiënten met een hoog risico.

Aangezien de kosten van de test vaak niet waren meegenomen in recent gepubliceerde kosteneffectiviteitsanalyses van geïndividualiseerde behandelstrategieën, was de analyse in eerste instantie beperkt tot een evaluatie van de verandering in behandeling. De minder intensieve behandeling voor patiënten met een laag risico leidde tot betere gezondheidsuitkomsten tegen lagere kosten. De nieuwe behandeling was dus dominant voor deze subgroep. Betere gezondheidsuitkomsten werden ook bereikt met de intensievere behandeling voor patiënten met een hoog risico, maar tegen substantieel hogere kosten. De incrementele kosteneffectiviteitsratio (IKER) van de verandering van behandeling voor die subgroep was €130,065 per gewonnen QALY.

Het includeren van de kosten van de test zorgden voor een stijging van de IKER voor beide behandelstrategieën. De kosteneffectiviteit wordt dus overschat indien de kosten van de test buiten beschouwing worden gelaten. De ernst van de overschatting is afhankelijk van de hoogte van de test, het verschil in gezondheidsuitkomsten als gevolg van de veranderde behandeling en de grootte van de subgroep voor wie de behandeling verandert. Op basis van deze bevindingen is een formule ontwikkeld waarmee de gevolgen van de test en behandeling apart kunnen worden geanalyseerd en gerapporteerd.

Aanbevelingen voor toekomstig onderzoek

Hoofdstuk 9 bespreekt de belangrijkste bevindingen en implicaties van dit proefschrift. Het is kosteneffectief om AML patiënten met een laag risico te behandelen met een minder intensieve behandeling. De nieuwe behandeling voor patiënten met een hoog risico, de allogene stamceltransplantatie van een ongerelateerde donor, is erg duur en heeft negatieve gevolgen voor de kwaliteit van leven. Kosteneffectievere behandelingen zijn daarom noodzakelijk voor deze subgroep. Verder kan de zorg voor AML verbeterd worden door meer aandacht te besteden aan ondersteunende zorg en het ontwikkelen van meer gerichte behandelingen voor andere subgroepen. Deze meer gerichte behandelingen kunnen mogelijk de behandelkosten verlagen indien een langdurige ziekenhuisopname niet meer noodzakelijk is.

Nieuwe klinische studies in AML moeten kwaliteit van leven meenemen als secondaire uitkomstmaat, omdat patiënten nog veel gezondheidsproblemen ervaren als gevolg van de ziekte of behandeling. Indien zowel de QLQ-C30 als de EQ-5D worden meegenomen

als kwaliteit van leven instrument, kan de validiteit van de EQ-5D bij acute leukemie ook nog verder worden onderzocht. Daarnaast moeten zowel directe als indirecte kosten worden gemeten in die studies om een adequaat oordeel te geven van de impact van nieuwe behandelstrategieën op de totale kosten.

De bevindingen uit dit proefschrift dragen ook bij aan de methodologie van kosteneffectiviteitsanalyses van geïndividualiseerde behandelstrategieën in andere ziektegebieden. De kosten van de test moeten altijd worden opgenomen in die analyses om een overschatting van de kosteneffectiviteit te voorkomen. Met behulp van de ontwikkelde formule kunnen de kosten van de test op een efficiënte wijze worden toegevoegd aan de analyses. De data verzameling en analyse kan nog steeds beperkt zijn tot de subgroep voor wie de behandeling verandert. Gestandaardiseerd gebruik van de formule zal ook de interpretatie van de resultaten verbeteren omdat vergelijkingen met andere studies beter te maken zijn. Tot slot kan de formule gebruikt worden in een vroeg stadium van productontwikkeling om richting te geven aan verder ontwikkeling en data verzameling.

LIST OF ABBREVIATIONS

AIC	Akaike information criteria
AML	Acute myeloid leukemia
APL	acute promyelocytic leukemia
*CEBPA*dm	*CCAAT enhancer binding protein* double mutations
CR	Complete remission
CS-PBM	Condition-specific preference-based measure
DAS28	Disease activity score
DES	discrete-event simulations
DFS	Disease-free survival
EDSS	Expended Disability Status Scale
EGFR	*epidermal growth factor receptor*
EQ-5D	EuroQol five-dimenisonal
EQ-5D-3L	Euroqol five dimensional - three level questionnaire
EQ-5D-5L	Euroqol five dimensional - five level questionnaire
EQ-VAS	EuroQol visual analogue scale
Erasmus MC	Erasmus Medical Center
ES	Effect size
ESR	Erythrocyte sedimentation rate
EVI1	*ecotropic virus integration-1*
FACT-B	Functional Assessment of Cancer Therapy-Breast
FLT3-ITD	Internal tandem duplication of the *fms-like tyrosine kinase-3 gene*
FN	False negative
FP	False positive
GVHD	graft-versus-host disease
HADS	Hospital-Anxiety and depressions scale
HAQ	Health assessment questionnaire
HDC	high-dose conditioning
HOVON	Hemato-oncology association for adults in the Nederlands
HR	Hazard ratio
HRQOL	Health-related quality of life
HSCT	Hematopoietic stem cell transplantation
HUI	Health utility index
ICER	Incremental cost-effectiveness ratio
IDH	*isocitrate dehydrogenase*
MAE	Mean absolute error
MS	Muliple sclerosis
MSIS-29	Multiple Sclerosis Impact Scale 29
MUD	Matched unrelated donor
NICE	National Institute of Health Care Excellence

NPM1	*nucleophosmin-1 gene mutations*
OMEP	orthogonal main effects plan
OS	Overall survival
PBM	Preference-based measure
PSA	Probabilistic sensitivity analysis
QALY	Quality-adjusted life year
QL scale	Global quality of life scale of the QLQ-C30
QLQ-C30	Quality of Life Questionnaire for Cancer 30
QLQ-PBM	Quality of Life Questionnaire - Preference Based Measure
RCT	Randomized controlled trial
REACH	Rotterdam Early Arthritis CoHort
RIC	Reduced-intensity conditioning
RMSE	Root mean square error
SD	Standard deviation
SF-36	Short-Form 36
TN	True negative
TP	True positive
TRM	Treatment-related mortality
TTO	Time trade-off
UCB	Umbilicial cord blood
WBC	White blood cell
WHO	World Health Organization

DANKWOORD

Promoveren is eigenlijk wel te vergelijken met een maken van een (hele grote) 'Wasgij' puzzel. Bij aanvang van mijn promotietraject had ik geen idee hoe mijn proefschrift er uiteindelijk uit zou gaan zien. Het enige dat ik wist was dat het 'iets' te maken zou hebben met de kosteneffectiviteit van nieuwe diagnostische mogelijkheden bij acute myeloïde leukemie. In de afgelopen jaren heb ik de puzzel beetje bij beetje opgelost en werd het eindresultaat steeds zichtbaarder. Op sommige momenten verliep het schrijven van mijn proefschrift heel soepel, maar er waren ook momenten dat het even duurde voordat ik een passend stukje vond. Gelukkig hoefde ik de puzzel niet helemaal alleen op te lossen en daarom wil ik in dit dankwoord iedereen bedanken die, op zijn of haar manier, een bijdrage heeft geleverd aan mijn proefschrift.

In de eerste plaats ben ik mijn promotoren, Carin Uyl-de Groot en Bob Löwenberg, en copromotor Ken Redekop dankbaar voor hun waardevolle begeleiding tijdens mijn promotietraject. Beste Carin, jouw passie om te zorgen voor een goede behandeling voor iedere patiënt is een voorbeeld voor mij. Ik hoop daar in de toekomst zeker nog meer aan bij te dragen. Ik ben je verder zeer dankbaar voor jouw steun om me verder te ontwikkelen op het gebied van onderwijs. Jouw betrokkenheid beperkte zich echter niet alleen tot het professionele vlak, maar je had ook altijd aandacht voor mij als mens. Helaas bleek er wel een selectie effect te zijn bij het lezen van de korfbaluitslagen in de krant, waardoor je mijn prestaties mogelijk onderschat hebt.

Beste Bob, ik vind het ontzettend fijn dat je bereid was om voor mij een tweede promotor te zijn. Ik heb van jou niet alleen veel geleerd over acute myeloïde leukemie, maar ook hoe ik mijn eigen onderzoeken moet positioneren in een groter geheel. Ik heb de samenwerking met jou als heel positief ervaren.

Beste Ken, de tijd die wij samen hebben besteed aan inhoudelijke discussies hebben mijn proefschrift aanzienlijk verbeterd. Het was heel fijn dat jouw deur altijd open stond en je altijd tijd vrij maakte om uitgebreid van gedachte te wisselen over de problemen waar ik tegen aanliep. Je hebt mij veel wijze lessen geleerd op het gebied van onderzoek en onderwijs. Ik bewonder jouw gave dat je altijd in staat bent om treffende beeldspraken te vinden voor welk onderwerp dan ook. Zelfs bij het doceren van gezondheidseconomie weet je een link te leggen met de serie 'The Big Bang Theory'. Je ziet dat ik in dit dankwoord ook een poging doe, maar ik kan jou nog lang niet evenaren.

De artikelen in mijn proefschrift waren niet tot stand gekomen zonder de hulp van de co-auteurs. In het bijzonder wil ik Matthijs bedanken voor de prettige samenwerking

tijdens onze kwaliteit van leven studies. Mijn artikelen zijn zeker ook verbeterd door vakinhoudelijke discussies met collega-onderzoekers op de werkvloer en bij congressen, zoals LoLaHESG en ISPOR.

Verder wil ik de promotiecommissie bedanken voor het kritisch doornemen van mijn proefschrift en het stellen van uitdagende vragen tijdens de verdediging.

Lieve Jennifer en Nicole, ik vind het heel fijn dat jullie naast mij staan tijdens de verdediging.

Jen, het grootste gedeelte van mijn promotietraject had ik de eer om een kamer met jou te delen. Ik heb genoten van onze tweetalige brainstormsessies over modellen, vroege HTA, artikelen, onderwijs, toekomstige carrièremogelijkheden en, misschien wel het belangrijkste, ons sociale leven. Jouw gedrevenheid voor het bereiken van je persoonlijke doelen zijn een voorbeeld voor mij!

Nicole, we hoefden al nooit te zoeken naar een gespreksonderwerp als we elkaar zagen, maar toen jij ook aan een promotietraject begon raakten we helemaal nooit meer uitgepraat. Heel fijn om een zo'n goede vriendin te hebben die precies weet hoe het is om te promoveren. Ik hoop dat we samen nog veel wijntjes zullen drinken op onze geaccepteerde artikelen!

Bij het leggen van een puzzel is het essentieel om soms even afstand te nemen en op een later moment weer verder te kijken. Op dat moment zie je ineens passende stukjes die je eerder over het hoofd zag. Gelukkig was de sfeer bij BMG zo goed dat hier ook de mogelijkheid toe was. Sociale hoogtepunten waren voor mij de vrijdagse kroket en deelname aan de Roparun. Buiten werktijd waren er ook meer dan genoeg mogelijkheden om de aandacht even ergens anders op te richten tijdens bijvoorbeeld het korfballen en gezellige etentjes of uitjes met familie en vrienden.

Lieve papa en mama, het klinkt heel cliché om te zeggen dat ik zonder jullie dit proefschrift niet had geschreven, maar het is echt waar. Papa, van jou heb ik mijn aanleg voor statistiek meegekregen terwijl ik de passie voor de gezondheidszorg van mama heb geërfd. Ik vind het dan ook extra bijzonder dat mijn proefschrift zo duidelijk een product is van jullie twee werelden. Verder was het ook heel fijn om op zondag te kunnen ontspannen tijdens een heerlijk diner in Trattoria Toscane. Lieve Carli en Lennert, die etentjes zijn een stuk gezelliger met jullie erbij. Bij deze wel het verzoek om me voortaan bij een plagerij over mijn lengte aan te spreken met mijn nieuwe titel!

Lieve Peter, dankjewel dat je mij hebt gesteund door mij de vrijheid te geven om mijn eigen doelen na te streven, maar er wel altijd voor mij te zijn bij een tegenslag. Jij tovert altijd weer een lach op mijn gezicht!

PhD PORTFOLIO

PhD student:	Annemieke Leunis
Institute:	institute of Health Policy & Management, Erasmus University Rotterdam, the Netherlands
PhD period:	2008-2015
Promotors:	Prof. dr. Carin A. Uyl-de Groot, Prof. dr. Bob Löwenberg
Co-promotor:	W. Ken Redekop

PhD training

Diagnostic Research – Netherlands Institute for Health Sciences (2008)

Advanced Diagnostic Research – Netherlands Institute for Health Sciences (2008)

Advanced Modelling Methods for Health Economic Evaluation Course – Centre for Health Economics, The University York (2008)

Presentation skills - institute of Health Policy & Management (2009)

Survival analysis – Netherlands Institute for Health Sciences (2009)

Discrete Event Simulation for Economic Analyses – International Society for Pharmaco-economics and Outcomes Research (2009)

Rasch analysis - The Psychometric Laboratory for Health Sciences Within the Academic Department of Rehabilitation Medicine (2009)

Academic writing in English for PhD students, Language & training centre, Erasmus University Rotterdam(2010)

Ready in four years (2010)

Basis course didactics - Risbo, Erasmus University Rotterdam (2012)

Supervising theses (2012-2013)

Teaching module 'Examination I' (2013)

Teaching module 'Plenair meetings' (2014)

Teaching module 'Examination II' (2014)

Teaching module 'Course design' (2014)

Presentations at (inter)national conferences

Podium presentations

Potential Health and Economic Impact of Genomics and Proteomics technology in acute myeloid leukemia, at the 4th Dutch Hematology Congress, Arnhem the Netherlands (2010)

The development and validation of a decision model representing the full disease course of acute myeloid leukemia, at the International Society for Pharmacoeconomics and Outcomes Research 13th Annual European Congress , Prague, Czech Republic (2010)

The direct treatment costs for patients with acute myeloid leukemia, at the 6th Dutch Hematology Congress, Arnhem, the Netherlands (2012)

The calculation of quality of life utilities for acute leukemia: a comparison between EQ5D-5L and QLQ-C30, at the International Society for Pharmacoeconomics and Outcomes Research 15th Annual European Congress, Berlin, Germany (2012)

Workshops

Issues in modelling the prognosis and treatment of non-communicable diseases using patient-level simulation: do the guidelines help us, at the International Society for Pharmacoeconomics and Outcomes Research 15th Annual European Congress, Berlin, Germany (2012)

Poster presentations

Challenges in using the literature to estimate the outcomes of current risk stratification methods in adult patients with primary acute myeloid leukemia, at the the International Society for Pharmacoeconomics and Outcomes Research 12th Annual European Congress, Paris, France (2009)

The cost-effectiveness of new diagnostic tests which identify prognostic subtypes in acute myeloid leukemia, at the 16th Congress of the European Hematology Association, London, the United Kingdom (2011)

Impaired health-related quality of life in acute myeloid leukemia survivors, at the 18th Congress of the European Hematology Association, Stockholm, Sweden (2013)

An efficient design for cost-effectiveness studies of personalized medicine strategies, at the International Society for Pharmacoeconomics and Outcomes Research 17th Annual European Congress, Amsterdam, the Netherlands (2014)

Other meetings, workshops and contributions

Annual meetings of the Center of Translational Molecular Medicine (2008 – 2013)

Research seminars at the institute of Health Policy & Management (2008-2015)

Attendence at the Lowlands Health Economists' Study Group (LolaHESG): papers discussed by other researchers and discussant of other researchers' papers (2009-2014)

Day-course Personalized medicine & Companion Diagnostics (2011)

CTMM course on ethical and societal issues (2012)

Early MTA and making decisions for new medical devices (2013)

Statistics book club (2013)

Teaching activities

Working groups Quantitative research methods, Bachelor 2 Health Sciences (2008-2009)

Computer practicum Public Health, Master Health Economics, Policy & Law (2008-2009)

Working groups and computer practicum Health Technology Assessment, Master Health Economics, Policy & Law (2008-2014)

Working groups Statistics, Bachelor 1 Health Sciences (2009-2010)

Co-coördination, lectures and working groups in the minor, the future of health care, Bachelor 3 Health Sciences (2011-2013)

Lecture and working groups Participating HTA, Master Health Economics, Policy & Law Master (2013-2015)

Coördinating, lectures and working groups Introduction in Health Sciences, Bachelor 1 Health Sciences (2013-2015)

Mentoring first year students (2014-2015)

Supervising and coevaluation several bachelor and master theses (2010-2015)

Other activities

Board member of Phd Council jBMG (2009-2014)

Scientific publications not included in this thesis

Leunis, A. & Varkevisser M. 2010. 'Internal medicine residents' perception of the learning environment in Dutch teaching hospitals.' Medical Teacher 32(1):93.

Scientific awards

Best new investigator podium presentation for the presentation 'The development and validation of a decision model representing the full disease course of acute myeloid leukemia' at the International Society for Pharmacoeconomics and Outcomes Research 13th Annual European Congress , Prague, Czech Republic

Nominated for the MTA prijs 2012 of the Dutch Society of Health Technology Assessment for the paper 'The development and validation of a decision-analytic model representing the full disease course of acute myeloid leukemia' published in Pharmacoeconomics 2012.

ABOUT THE AUTHOR

Annemieke Leunis was born in Strijen on the 17th of September 1986. From 2004 to 2008 she studied Health Sciences at the Erasmus University and obtained her master's degree cum laude in health economics. During her study, she worked as a student assistant and taught working. After graduation, she started as a junior researcher at institute of Medical Technology Assessment of the institute of Health Policy and Management at the Erasmus University in Rotterdam. She has broad research interests in the field of health economics, including the measurement of health care costs and quality of life as well as modeling for cost-effectiveness analyses. Besides her PhD project, she has worked on several cost-effectiveness analyses of new drugs. In addition, she taught various bachelor and master courses at the institute of Health Policy and Management, including introduction to health care sciences, statistics and health technology assessment. She has also supervised bachelor and master theses.

REFERENCES

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012. Eur J Cancer. 2013;49(6):1374-1403.
2. OECD. Health at a glance: Europe 2012. 2012.
3. Berrino F, De Angelis R, Sant M, et al. Survival for eight major cancers and all cancers combined for european adults diagnosed in 1995–99: Results of the EUROCARE-4 study. The Lancet Oncology. 2007;8(9):773-783.
4. Karim-Kos HE, de Vries E, Soerjomataram I, Lemmens V, Siesling S, Coebergh JWW. Recent trends of cancer in europe: A combined approach of incidence, survival and mortality for 17 cancer sites since the 1990s. Eur J Cancer. 2008;44(10):1345-1389.
5. Verdecchia A, Francisci S, Brenner H, et al. Recent cancer survival in europe: A 2000–02 period analysis of EUROCARE-4 data. The Lancet Oncology. 2007;8(9):784-796.
6. Luengo-Fernandez R, Leal J, Gray A, Sullivan R. Economic burden of cancer across the european union: A population-based cost analysis. The Lancet Oncology. 2013;14(12):1165-1174.
7. Mariotto AB, Robin Yabroff K, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the united states: 2010–2020. Journal of the National Cancer Institute. 2011;103(2):117-128.
8. Garattini S, Bertele V. Efficacy, safety, and cost of new anticancer drugs. BMJ. 2002;325(7358): 269-271.
9. Roberts TG, Chabner BA. Beyond fast track for drug approvals. N Engl J Med. 2004;351(5):501-505.
10. Sears C, Armstrong SA. Microarrays to identify new therapeutic strategies for cancer. Adv Cancer Res. 2007;96:51-74.
11. Schleidgen S, Klingler C, Bertram T, Rogowski WH, Marckmann G. What is personalized medicine: Sharpening a vague term based on a systematic literature review. BMC Med Ethics. 2013;14:55-6939-14-55.
12. Redekop WK, Mladsi D. The faces of personalized medicine: A framework for understanding its meaning and scope. Value Health. 2013;16(6 Suppl):S4-9.
13. Appelbaum FR, Gundacker H, Head DR, et al. Age and acute myeloid leukemia. Blood. 2006; 107(9):3481-3485.
14. Haynes AB, You YN, Hu C, et al. Postoperative chemotherapy use after neoadjuvant chemoradiotherapy for rectal cancer. Cancer. 2014:n/a-n/a.
15. Steele SR, Park GE, Johnson EK, et al. The impact of age on colorectal cancer incidence, treatment, and outcomes in an equal-access health care system. Dis Colon Rectum. 2014;57(3):303-310.
16. Geraci JM, Escalante CP, Freeman JL, Goodwin JS. Comorbid disease and cancer: The need for more relevant conceptual models in health services research. Journal of Clinical Oncology. 2005; 23(30):7399-7404.
17. Sogaard M, Thomsen RW, Bossen KS, Sorensen HT, Norgaard M. The impact of comorbidity on cancer survival: A review. Clin Epidemiol. 2013;5(Suppl 1):3-29.
18. Aleksandrova K, Pischon T, Jenab M, et al. Combined impact of healthy lifestyle factors on colorectal cancer: A large european cohort study. BMC Med. 2014;12(1):168.
19. Balmaña J, Diez O, Rubio I, Castiglione M, On behalf of the ESMO Guidelines Working Group. BRCA in breast cancer: ESMO clinical practice guidelines. Annals of Oncology. 2010;21(suppl 5):v20-v22.
20. Viani GA, Bernardes da Silva LG, Stefano EJ. Prognostic indexes for brain metastases: Which is the most powerful? International Journal of Radiation Oncology*Biology*Physics. 2012;83(3): e325-e330.

21. Löwenberg B. Acute myeloid leukemia: The challenge of capturing disease variety. ASH Education Program Book. 2008;2008(1):1-11.
22. Aparicio T, Jouve J, Teillet L, et al. Geriatric factors predict chemotherapy feasibility: Ancillary results of FFCD 2001-02 phase III study in first-line chemotherapy for metastatic colorectal cancer in elderly patients. Journal of Clinical Oncology. 2013;31(11):1464-1470.
23. Paz-Ares L, Soulières D, Melezínek I, et al. Clinical outcomes in non-small-cell lung cancer patients with EGFR mutations: Pooled analysis. J Cell Mol Med. 2010;14(1-2):51-69.
24. Lostumbo L, Carbine NE, Wallace J. Prophylactic mastectomy for the prevention of breast cancer. Cochrane Database Syst Rev. 2010;(11):CD002748. doi(11):CD002748.
25. Carey LA, Perou CM, Livasy CA,et al. RAce, breast cancer subtypes, and survival in the carolina breast cancer study. JAMA. 2006;295(21):2492-2502.
26. Dohner H, Estey EH, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: Recommendations from an international expert panel, on behalf of the european LeukemiaNet. Blood. 2010;115(3):453-474.
27. Nagaraja, V., Eslick,G. Forthcoming prognostic markers for esophageal cancer: A systematic review and meta-analysis. Journal of Gastrointestinal Oncology. 2014;5(1).
28. Wong A, Wouterse B, Slobbe LCJ, Boshuizen HC, Polder JJ. Medical innovation and age-specific trends in health care utilization: Findings and implications. Soc Sci Med. 2012;74(2):263-272.
29. Drummond M, Sculpher MJ, Stoddart GL, Torrance GW, O'Brien BJ. Methods for the economic evaluation of health care. 3rd edition ed. Oxford University Press; 2005.
30. National institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal 2013. London: NICE2013.
31. Whitehurst DGT, Bryan S, Lewis M. Systematic review and empirical comparison of contemporaneous EQ-5D and SF-6D group mean scores. Medical Decision Making. 2011;31(6):E34-E44.
32. Brazier J, Yang Y, Tsuchiya A, Rowen D. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. The European Journal of Health Economics. 2010;11(2):215-225.
33. Brazier JE, Rowen D, Mavranezouli I, et al. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess. 2012;16(32):1-114.
34. Brazier JE, Longworth L. NICE DSU technical support document 8: An introduction to the measurement and valuation health for NICE submissions. Available from http://www.nicedsu.org.uk. Updated 2011. [Accessed April, 29, 2013].
35. Hoefman RJ, van Exel J, Brouwer W. How to include informal care in economic evaluations. Pharmacoeconomics. 2013;31(12):1105-1119.
36. Krol M, Brouwer W, Rutten F. Productivity costs in economic evaluations: Past, present, future. Pharmacoeconomics. 2013;31(7):537-549.
37. Veenstra DL, Roth JA, Garrison LP,Jr, Ramsey SD, Burke W. A formal risk-benefit framework for genomic tests: Facilitating the appropriate translation of genomics into clinical practice. Genet Med. 2010;12(11):686-693.
38. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices?modeling studies. Value in Health. 2003;6(1):9-17.
39. Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. New York: Oxford University Press; 2006.

40. Claxton KP, Sculpher MJ. Using value of information analysis to prioritise health research: Some lessons from recent UK experience. Pharmacoeconomics. 2006;24(11):1055-1068.
41. Annemans L, Geneste B, Jolain B. Early modelling for assessing health and economic outcomes of drug therapy. Value Health. 2000;3(6):427-434.
42. Estey EH. Acute myeloid leukemia: 2012 update on diagnosis, risk stratification, and management. Am J Hematol. 2012;87(1):89-99.
43. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the world health organization (WHO) classification of myeloid neoplasms and acute leukemia: Rationale and important changes. Blood. 2009;114(5):937-951.
44. Byrd JC, Mrozek K, Dodge RK, et al. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: Results from cancer and leukemia group B (CALGB 8461). Blood. 2002;100(13):4325-4336.
45. Grimwade D, Walker H, Harrison G, et al. The predictive value of hierarchical cytogenetic classification in older adults with acute myeloid leukemia (AML): Analysis of 1065 patients entered into the united kingdom medical research council AML11 trial. Blood. 2001;98(5):1312-1320.
46. Slovak ML, Kopecky KJ, Cassileth PA, et al. Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: A southwest oncology group/eastern cooperative oncology group study. Blood. 2000;96(13):4075-4083.
47. Martelli MP, Sportoletti P, Tiacci E, Martelli MF, Falini B. Mutational landscape of AML with normal cytogenetics: Biological and clinical implications. Blood Rev. 2013;27(1):13-22.
48. Sanz MA. Treatment of acute promyelocytic leukemia. ASH Education Program Book. 2006; 2006(1):147-155.
49. Lo-Coco F, Ammatuna E, Montesinos P, Sanz MA. Acute promyelocytic leukemia: Recent advances in diagnosis and management. Semin Oncol. 2008;35(4):401-409.
50. Estey EH. Acute myeloid leukemia: 2013 update on risk-stratification and management. Am J Hematol. 2013;88(4):317-327.
51. Hatzimichael E, Georgiou G, Benetatos L, Briasoulis E. Gene mutations and molecularly targeted therapies in acute myeloid leukemia. Am J Blood Res. 2013;3(1):29-51.
52. Konig H, Levis M. Is targeted therapy feasible in acute myelogenous leukemia? Current Hematologic Malignancy Reports. 2014:1-10.
53. Redaelli A, Stephens JM, Brandt S, Botteman MF, Pashos CL. Short- and long-term effects of acute myeloid leukemia on patient health-related quality of life. Cancer Treat Rev. 2004;30(1):103-117.
54. Zittoun R, Suciu S, Watson M, et al. Quality of life in patients with acute myelogenous leukemia in prolonged first complete remission after bone marrow transplantation (allogeneic or autologous) or chemotherapy: A cross-sectional study of the EORTC-GIMEMA AML 8A trial. Bone Marrow Transplant. 1997;20(4):307-315.
55. Watson M, Buck G, Wheatley K, et al. Adverse impact of bone marrow transplantation on quality of life in acute myeloid leukaemia patients: Analysis of the UK medical research council AML 10 trial. Eur J Cancer. 2004;40(7):971-978.
56. Messerer D, Engel J, Hasford J, et al. Impact of different post-remission strategies on quality of life in patients with acute myeloid leukemia. Haematologica. 2008;93(6):826-833.
57. Maynadié M, De Angelis R, Marcos-Gragera R, et al. Survival of european patients diagnosed with myeloid malignancies: A HAEMACARE study. Haematologica. 2013;98(2):230-238.
58. Sant M, Allemani C, Tereanu C, et al. Incidence of hematologic malignancies in europe by morphologic subtype: Results of the HAEMACARE project. Blood. 2010;116(19):3724-3734.

59. Rowe JM. Optimal induction and post-remission therapy for AML in first remission. ASH Education Program Book. 2009;2009(1):396-405.
60. OECD. Health at a glance 2011. Organisation for Economic Co-operation and Development; 2011. 10.1787/health_glance-2011-en.
61. Stalfelt AM, Brodin H. Costs over time in conventional treatment of acute myeloid leukaemia. A study exploring changes in treatment strategies over two decades. J Intern Med. 1994;236(4): 401-409.
62. Uyl-de Groot CA, Lowenberg B, Vellenga E, Suciu S, Willemze R, Rutten FF. Cost-effectiveness and quality-of-life assessment of GM-CSF as an adjunct to intensive remission induction chemotherapy in elderly patients with acute myeloid leukemia. Br J Haematol. 1998;100(4):629-636.
63. Uyl-de Groot CA, Gelderblom-den Hartog J, Huijgens PC, Willemze R, van Ineveld BM. Costs of diagnosis, treatment, and follow up of patients with acute myeloid leukemia in the netherlands. J Hematother Stem Cell Res. 2001;10(1):187-192.
64. Burnett A, Wetzler M, Löwenberg B. Therapeutic advances in acute myeloid leukemia. Journal of Clinical Oncology. 2011;29(5):487-494.
65. Cornelissen JJ, Löwenberg B. Role of allogeneic stem cell transplantation in current treatment of acute myeloid leukemia. ASH Education Program Book. 2005;2005(1):151-155.
66. Estey E, de Lima M, Tibes R, et al. Prospective feasibility analysis of reduced-intensity conditioning (RIC) regimens for hematopoietic stem cell transplantation (HSCT) in elderly patients with acute myeloid leukemia (AML) and high-risk myelodysplastic syndrome (MDS). Blood. 2007;109(4): 1395-1400.
67. King ME, Rowe JM. Recent developments in acute myelogenous leukemia therapy. Oncologist. 2007;12 Suppl 2:14-21.
68. The Dutch Healthcare Authority. DBC-tariefapplicatie. Available from http://dbc-tarieven.nza.nl/Nzatarieven/top.do. [Accessed 04/29, 2011].
69. Hakkaart-van Roijen L, Tan SS, Bouwmans CA. Handleiding voor kostenonderzoek. methoden en standaard kostprijzen voor economische evaluaties in de gezondheidszorg. Diemen: CVZ2010.
70. Franken MG, Gaultney JG, Blommestein HM, et al. Pilot outcomes research: Effects and costs of bortezomib in relapsed or refractory multiple myeloma. Diemen: College voor zorgverzekeringen 2012. ; No. 2012075715. http://www.cvz.nl/binaries/content/documents/cvzinternet/nl/documenten/rubriek+zorgpakket/imta-onderzoek-dure-geneesmiddelen-1207.pdf.
71. Blommestein HM, Verelst SG, Huijgens PC, Blijlevens NM, Cornelissen JJ, Uyl-de Groot CA. Real-world costs of autologous and allogeneic stem cell transplantations for haematological diseases: A multicentre study. Ann Hematol. 2012;91(12):1945-1952.
72. Statistics Netherlands. Statline. Available from http://statline.cbs.nl/StatWeb/?LA=en. [Accessed 04/29, 2011].
73. Solomon SR, Matthews RH, Barreras AM, et al. Outpatient myeloablative allo-SCT: A comprehensive approach yields decreased hospital utilization and low TRM. Bone Marrow Transplant. 2010; 45(3):468-475.
74. Riley GF, Lubitz JD. Long-term trends in medicare payments in the last year of life. Health Serv Res. 2010;45(2):565-576.
75. Stalfelt AM, Brodin H, Wadman B. Cost analysis of different phases of acute myeloid leukaemia. Leuk Res. 1994;18(10):783-790.
76. Nerich V, Lioure B, Rave M, et al. Induction-related cost of patients with acute myeloid leukaemia in france. Int J Clin Pharm. 2011;33(2):191-199.

77. Cordonnier C, Maury S, Esperou H, et al. Do minitransplants have minicosts? A cost comparison between myeloablative and nonmyeloablative allogeneic stem cell transplant in patients with acute myeloid leukemia. Bone Marrow Transplant. 2005;36(7):649-654.
78. Saito AM, Zahrieh D, Cutler C, et al. Lower costs associated with hematopoietic cell transplantation using reduced intensity vs high-dose regimens for hematological malignancy. Bone Marrow Transplant. 2007;40(3):209-217.
79. Rosenblat TL, Jurcic JG. Induction and postremission strategies in acute myeloid leukemia: State of the art and future directions. Hematol Oncol Clin North Am. 2011;25(6):1189-1213.
80. Moller T, Nielsen OJ, Welinder P, et al. Safe and feasible outpatient treatment following induction and consolidation chemotherapy for patients with acute leukaemia. Eur J Haematol. 2010;84(4): 316-322.
81. Walter RB, Lee SJ, Gardner KM, et al. Outpatient management following intensive induction chemotherapy for myelodysplastic syndromes and acute myeloid leukemia: A pilot study. Haematologica. 2011;96(6):914-917.
82. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ. 1986; 5(1):1-30.
83. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271-292.
84. Petrillo J, Cairns J. Converting condition-specific measures into preference-based outcomes for use in economic evaluation. Expert Rev Pharmacoecon Outcomes Res. 2008;8(5):453-461.
85. Versteegh MM, Rowen D, Brazier JE, Stolk EA. Mapping onto eq-5 D for patients in poor health. Health Qual Life Outcomes. 2010;8:141-7525-8-141.
86. Williams A. The EuroQol instrument. In: Kind P, Brooks R, Rabin R, eds. EQ-5D concepts and methods: A developmental history. Dordrecht: Springer; 2005.
87. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The dutch tariff: Results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10): 1121-1132.
88. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-1108.
89. Aaronson NK, Ahmedzai S, Bergman B, 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 Natl Cancer Inst. 1993;85(5):365-376.
90. McKenzie L, van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: The potential to estimate QALYs without generic preference data. Value Health. 2008.
91. Bruce B, Fries JF. The stanford health assessment questionnaire: A review of its history, issues, progress, and documentation. J Rheumatol. 2003;30(1):167-178.
92. Bansback N, Marra C, Tsuchiya A, et al. Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. Arthritis Rheum. 2007; 57(6):963-971.
93. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The multiple sclerosis impact scale (MSIS-29): A new patient-based outcome measure. Brain. 2001;124(Pt 5):962-973.
94. Segeren CM, Sonneveld P, van der Holt B, et al. Overall and event-free survival are not improved by the use of myeloablative therapy following intensified chemotherapy in previously untreated patients with multiple myeloma: A prospective randomized phase 3 study. Blood. 2003;101(6): 2144-2151.

95. Doorduijn JK, van der Holt B, van Imhoff GW, et al. CHOP compared with CHOP plus granulocyte colony-stimulating factor in elderly patients with aggressive non-hodgkin's lymphoma. J Clin Oncol. 2003;21(16):3041-3050.
96. Boggild M, Palace J, Barton P, et al. Multiple sclerosis risk sharing scheme: Two year results of clinical cohort study with historical comparator. BMJ. 2009;339:b4677.
97. Tsuchiya A, Brazier J, McColl E, Parkin D. Deriving preference-based single indices from non-preference based condition-specific instruments: Converting AQLQ into EQ5D indices. Sheffield: ScHARR, Sheffield Health Economics Group, University of Sheffield, UK2002. ; No. Report No.: Discussion Paper Series 02/1.
98. Szende A, Oppe M, Devlin N, eds. EQ-5D value sets. inventory, comparative review and user guide. Dordrecht: Springer; 2007.
99. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. New York: Oxford University Press; 2007.
100. Kontodimopoulos N, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value Health. 2009;12(8): 1151-1157.
101. Ramp M, Khan F, Misajon RA, Pallant JF. Rasch analysis of the multiple sclerosis impact scale MSIS-29. Health Qual Life Outcomes. 2009;7:58-7525-7-58.
102. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002;40(2):113-128.
103. Yang Y, Brazier JE, Tsuchiya A, Young TA. Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. Med Decis Making. 2011;31(2):281-291.
104. Brazier J, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C. Estimation of a preference-based index from a condition-specific measure: The king's health questionnaire. Med Decis Making. 2008;28(1):113-126.
105. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: What happens to cross programme comparability? Health Econ. 2010;19(2):125-129.
106. Fryback DG, Lawrence WF,Jr. Dollars may not buy as many QALYs as we think: A problem with defining quality-of-life adjustments. Med Decis Making. 1997;17(3):276-284.
107. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004;13(9):873-884.
108. Kind P, Brooks R, Rabin R, eds. EQ-5D concepts and methods: A developmental history. Dordrecht: Springer; 2005.
109. ten Klooster PM, Taal E, van de Laar MA. Rasch analysis of the dutch health assessment questionnaire disability index and the health assessment questionnaire II in patients with rheumatoid arthritis. Arthritis Rheum. 2008;59(12):1721-1728.
110. Versteegh MM, Leunis A, Luime JJ, Boggild M, Uyl-de Groot CA, Stolk EA. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. Med Decis Making. 2012;32(4):554-568.
111. Young TA, Yang Y, Brazier JE, Tsuchiya A. The use of rasch analysis in reducing a large condition-specific instrument for preference valuation: The case of moving from AQLQ to AQL-5D. Med Decis Making. 2011;31(1):195-210.
112. Mavranezouli I, Brazier JE, Young TA, Barkham M. Using rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). Qual Life Res. 2011;20(3):321-333.

113. Pallant JF, Tennant A. An introduction to the rasch measurement model: An example using the hospital anxiety and depression scale (HADS). Br J Clin Psychol. 2007;46(Pt 1):1-18.
114. Tennant A, McKenna SP, Hagell P. Application of rasch analysis in the development and application of quality of life instruments. Value Health. 2004;7 Suppl 1:S22-6.
115. Brazier JE, Roberts J, Platts M, Zoellner YF. Estimating a preference-based index for a menopause specific health quality of life questionnaire. Health Qual Life Outcomes. 2005;3:13.
116. Stolk EA, Busschbach JJ. Validity and feasibility of the use of condition-specific outcome measures in economic evaluation. Qual Life Res. 2003;12(4):363-371.
117. Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. Eur J Health Econ. 2010;11(4):427-434.
118. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: A revised version of the health assessment questionnaire. Arthritis Rheum. 2004;50(10):3296-3305.
119. Derolf ÅR, Kristinsson SY, Andersson TM-, Landgren O, Dickman PW, Björkholm M. Improved patient survival for acute myeloid leukemia: A population-based study of 9729 patients diagnosed in sweden between 1973 and 2005. Blood. 2009;113(16):3666-3672.
120. Pulte D, Gondos A, Brenner H. Improvements in survival of adults diagnosed with acute myeloblastic leukemia in the early 21st century. Haematologica. 2008;93(4):594-600.
121. Pulte D, Gondos A, Brenner H. Improvement in survival in younger patients with acute lymphoblastic leukemia from the 1980s to the early 21st century. Blood. 2009;113(7):1408-1411.
122. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the united states. Quality of Life Research. 2000;9(8):887-900.
123. Bevans M. Health-related quality of life following allogeneic hematopoietic stem cell transplantation. ASH Education Program Book. 2010;2010(1):248-254.
124. Efficace F, Novik A, Vignetti M, Mandelli F, Cleeland CS. Health-related quality of life and symptom assessment in clinical research of patients with hematologic malignancies: Where are we now and where do we go from here? Haematologica. 2007;92(12):1596-1598.
125. Efficace F, Kemmler G, Vignetti M, Mandelli F, Molica S, Holzner B. Health-related quality of life assessment and reported outcomes in leukaemia randomised controlled trials – A systematic review to evaluate the added value in supporting clinical decision making. Eur J Cancer. 2008; 44(11):1497-1506.
126. Syrjala KL, Langer SL, Abrams JR, Storer BE, Martin PJ. Late effects of hematopoietic cell transplantation among 10-year adult survivors compared with case-matched controls. Journal of Clinical Oncology. 2005;23(27):6596-6606.
127. Löwenberg B, van Putten W, Theobald M, et al. Effect of priming with granulocyte colony-stimulating factor on the outcome of chemotherapy for acute myeloid leukemia. N Engl J Med. 2003;349(8):743-752.
128. Löwenberg B, Pabst T, Vellenga E, et al. Cytarabine dose for acute myeloid leukemia. N Engl J Med. 2011;364(11):1027-1036.
129. Löwenberg B, Ossenkoppele GJ, van Putten W, et al. High-dose daunorubicin in older patients with acute myeloid leukemia. N Engl J Med. 2009;361(13):1235-1248.
130. Ossenkoppele GJ, Stussi G, Maertens J, et al. Addition of bevacizumab to chemotherapy in acute myeloid leukemia at older age: A randomized phase 2 trial of the dutch-belgian cooperative trial group for hemato-oncology (HOVON) and the swiss group for clinical cancer research (SAKK). Blood. 2012;120(24):4706-4711.

131. HOVON – the Haemato Oncology Foundation for Adults in the Netherlands. Available from http://www.hovon.nl/trials/trials-by-type/aml.html. Updated 2011. [Accessed 03/22, 2013].
132. Brooks R. EuroQol: The current state of play. Health Policy. 1996;37(1):53-72.
133. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20:1727-1736.
134. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007;45(3):259-263.
135. van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15(5):708-715.
136. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996;5(6):555-567.
137. Konig HH, Bernert S, Angermeyer MC, et al. Comparison of population health status in six european countries: Results of a representative survey using the EQ-5D questionnaire. Med Care. 2009;47(2):255-261.
138. van de Poll-Franse LV, Mols F, Gundy CM, et al. Normative data for the EORTC QLQ-C30 and EORTC-sexuality items in the general dutch population. Eur J Cancer. 2011;47(5):667-675.
139. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273(1):59-65.
140. Holland JC. History of psycho-oncology: Overcoming attitudinal and conceptual barriers. Psychosomatic Medicine. 2002;64(2):206-221.
141. Cocks K, King MT, Velikova G, Fayers PM, Brown JM. Quality, interpretation and presentation of european organisation for research and treatment of cancer quality of life questionnaire core 30 data in randomised controlled trials. Eur J Cancer. 2008;44(13):1793-1798.
142. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol. 2004;57(11):1138-1146.
143. Daniëls L, Oerlemans S, Krol AG, Poll-Franse L, Creutzberg C. Persisting fatigue in hodgkin lymphoma survivors: A systematic review. Ann Hematol. 2013;92(8):1023-1032.
144. Minton O, Stone P. How common is fatigue in disease-free breast cancer survivors? A systematic review of the literature. Breast Cancer Res Treat. 2008;112(1):5-13.
145. Persoon S, Kersten MJ, van der Weiden K, et al. Effects of exercise in patients treated with stem cell transplantation for a hematologic malignancy: A systematic review and meta-analysis. Cancer Treat Rev. 2013;39(6):682-690.
146. Puetz TW, Herring MP. Differential effects of exercise on cancer-related fatigue during and following treatment: A meta-analysis. Am J Prev Med. 2012;43(2):e1-e24.
147. Eom CS, Shin DW, Kim SY, et al. Impact of perceived social support on the mental health and health-related quality of life in cancer patients: Results from a nationwide, multicenter survey in south korea. Psychooncology. 2012;22:1283-1290.
148. Luszczynska A, Pawlowska I, Cieslak R, Knoll N, Scholz U. Social support and quality of life among lung cancer patients: A systematic review. Psychooncology. 2012;22:2160-2168.
149. Waters EA, Liu Y, Schootman M, Jeffe DB. Worry about cancer progression and low perceived social support: Implications for quality of life among early-stage breast cancer patients. Ann Behav Med. 2012;45:57-68.
150. Pidala J, Anasetti C, Jim H. Quality of life after allogeneic hematopoietic cell transplantation. Blood. 2009;114(1):7-19.

151. Koreth J, Schlenk R, Kopecky KJ, et al. Allogeneic stem cell transplantation for acute myeloid leukemia in first complete remission. JAMA: The Journal of the American Medical Association. 2009;301(22):2349-2361.
152. Baker F, Denniston M, Smith T, West MM. Adult cancer survivors: How are they faring? Cancer. 2005;104(S11):2565-2576.
153. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: A theoretical model. Soc Sci Med. 1999;48(11):1507-1515.
154. Sherman AC, Simonton S, Latif U, Plante TG, Anaissie EJ. Changes in quality-of-life and psychosocial adjustment among multiple myeloma patients treated with high-dose melphalan and autologous stem cell transplantation. Biol Blood Marrow Transplant. 2009;15(1):12-20.
155. Gottlieb BH, Bergen AE. Social support concepts and measures. J Psychosom Res. 2010;69(5): 511-520.
156. Pickard AS, Ray S, Ganguli A, Cella D. Comparison of FACT- and EQ-5D-based utility scores in cancer. Value Health. 2012;15(2):305-311.
157. Lamers LM, Bouwmans CAM, van Straten A, Donker MCH, Hakkaart L. Comparison of EQ-5D and SF-6D utilities in mental health patients. Health Econ. 2006;15(11):1229-1236.
158. Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA. Condition-specific preference-based measures: Benefit or burden? Value Health. 2012;15(3):504-513.
159. Rowen D, Young T, Brazier J, Gaugris S. Comparison of generic, condition-specific, and mapped health state utility values for multiple myeloma cancer. Value in Health. 2012;15(8):1059-1068.
160. Kontodimopoulos N, Pappa E, Papadopoulos A, Tountas Y, Niakas D. Comparing SF-6D and EQ-5D utilities across groups differing in health status. Quality of Life Research. 2009;18(1):87-97.
161. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. 2013. Available at http://publications.nice.org.uk/pmg9. [Accessed January 30, 2015].
162. Conner-Spady B, Cumming C, Nabholtz JM, Jacobs P, Stewart D. Responsiveness of the EuroQol in breast cancer patients undergoing high dose chemotherapy. Qual Life Res. 2001;10(6):479-486.
163. Kim S, Hwang J, Kim T, Hong Y, Jo M. Validity and reliability of the EQ-5D for cancer patients in korea. Supportive Care in Cancer. 2012;20(12):3155-3160.
164. Krabbe PF, Peerenboom L, Langenhoff BS, Ruers TJ. Responsiveness of the generic EQ-5D summary measure compared to the disease-specific EORTC QLQ C-30. Qual Life Res. 2004;13(7): 1247-1253.
165. Kvam AK, Fayers PM, Wisloff F. Responsiveness and minimal important score differences in quality-of-life questionnaires: A comparison of the EORTC QLQ-C30 cancer-specific questionnaire to the generic utility questionnaires EQ-5D and 15D in patients with multiple myeloma. Eur J Haematol. 2011;87(4):330-337.
166. Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey. Health Technol Assess. 2014;18(9):1-224.
167. Teckle P, Peacock S, McTaggart-Cowan H, et al. The ability of cancer-specific and generic preference-based instruments to discriminate across clinical and self-reported measures of cancer severities. Health Qual Life Outcomes. 2011;9:106-7525-9-106.
168. Janssen MF, Pickard AS, Golicki D, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: A multi-country study. Quality of Life Research. 2012:1-11.
169. Kim S, Kim H, Lee S, Jo M. Comparing the psychometric properties of the EQ-5D-3L and EQ-5D-5L in cancer patients in korea. Quality of Life Research. 2012;21(6):1065-1073.

170. Scalone L, Ciampichini R, Fagiuoli S, et al. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. Quality of Life Research. 2012:1-10.
171. Lee CF, Luo N, Ng R, et al. Comparison of the measurement properties between a short and generic instrument, the 5-level EuroQoL group's 5-dimension (EQ-5D-5L) questionnaire, and a longer and disease-specific instrument, the functional assessment of cancer therapy-breast (FACT-B), in asian breast cancer patients. Qual Life Res. 2013;22(7):1745-1751.
172. Leunis A, Ken Redekop W, Uyl-de Groot CA, Lowenberg B. Impaired health-related quality of life in acute myeloid leukemia survivors: A single-center study. Eur J Haematol. 2014;93(3):198-206.
173. HOVON – the Haemato Oncology Foundation for Adults in the Netherlands. Available from http://www.hovon.nl/trials/trials-by-type/aml.html. [Accessed April, 29, 2013].
174. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New Jersey: Lawrence Elbaum Associates, Inc,. publishers; 1988.
175. Garau M, Shah K, Mason A, Wang Q, Towse A, Drummond M. Using QALYs in cancer. Pharmacoeconomics. 2011;29(8):673-685.
176. Yang Y, Rowen D, Brazier J, Tsuchiya A, Young T, Longworth L. An exploratory study to test the impact on three "bolt-on" items to the EQ-5D. Value Health. 2015;18(1):52-60.
177. Swinburn P, Lloyd A, Boye KS, Edson-Heredia E, Bowman L, Janssen B. Development of a disease-specific version of the EQ-5D-5L for use in patients suffering from psoriasis: Lessons learned from a feasibility study in the UK. Value in Health. 2013;16(8):1156-1162.
178. Yang Y, Brazier J, Tsuchiya A. Effect of adding a sleep dimension to the EQ-5D descriptive system: A "Bolt-on" experiment. Medical Decision Making. 2014;34(1):42-53.
179. Krabbe PFM, Stouthard MEA, Essink-Bot M, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. J Clin Epidemiol. 1999;52(4):293-301.
180. Rapkin B, Schwartz C. Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. Health and Quality of Life Outcomes. 2004;2(1):14.
181. Ahmed S, Ring L. Influence of response shift on evaluations of change in patient-reported outcomes. Expert Review of Pharmacoeconomics & Outcomes Research. 2008;8(5):479-489.
182. Ginsburg GS, Willard HF. Genomic and personalized medicine: Foundations and applications. Transl Res. 2009;154(6):277-287.
183. Salvesen HB, Haldorsen IS, Trovik J. Markers for individualised therapy in endometrial carcinoma. Lancet Oncol. 2012;13(8):e353-61.
184. Romano E, Schwartz GK, Chapman PB, Wolchock JD, Carvajal RD. Treatment implications of the emerging molecular classification system for melanoma. Lancet Oncol. 2011;12(9):913-922.
185. Marcucci G, Haferlach T, Dohner H. Molecular genetics of adult acute myeloid leukemia: Prognostic and therapeutic implications. J Clin Oncol. 2011;29(5):475-486.
186. Brenton JD, Carey LA, Ahmed AA, Caldas C. Molecular classification and molecular forecasting of breast cancer: Ready for clinical application? Journal of Clinical Oncology. October 10, 2005; 23(29):7350-7360.
187. Rubin MA, Maher CA, Chinnaiyan AM. Common gene rearrangements in prostate cancer. Journal of Clinical Oncology. 2011;29(27):3659-3668.
188. Albain KS, Paik S, van't Veer L. Prediction of adjuvant chemotherapy benefit in endocrine responsive, early breast cancer using multigene assays. Breast. 2009;18 Suppl 3:S141-5.
189. Cornelissen JJ, Van Putten WLJ, Verdonck LF, et al. Results of a HOVON/SAKK donor versus no-donor analysis of myeloablative HLA-identical sibling stem cell transplantation in first remission

acute myeloid leukemia in young and middle-aged adults: Benefits for whom? Blood. 2007; 109(9):3658-3666.
190. Schlenk RF, Döhner K, Krauter J, et al. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. N Engl J Med. 2008;358(18):1909-1918.
191. Wouters BJ, Löwenberg B, Erpelinck-Verschueren CAJ, van Putten WLJ, Valk PJM, Delwel R. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. Blood. 2009;113(13):3088-3091.
192. Sanderson S, Zimmern R, Kroese M, Higgins J, Patch C, Emery J. How can the evaluation of genetic tests be enhanced? lessons learned from the ACCE framework and evaluating genetic tests in the united kingdom. Genet Med. 2005;7(7):495-500.
193. Scott SA. Personalizing medicine with clinical pharmacogenetics. Genet Med. 2011;13(12): 987-995.
194. Frueh FW. Back to the future: Why randomized controlled trials cannot be the answer to pharmacogenomics and personalized medicine. Pharmacogenomics. 2009;10(7):1077-1081.
195. Al-Badriyeh D, Slavin M, Liew D, et al. Pharmacoeconomic evaluation of voriconazole versus posaconazole for antifungal prophylaxis in acute myeloid leukaemia. Journal of Antimicrobial Chemotherapy. 2010;65(5):1052-1061.
196. Kurosawa S, Yamaguchi T, Miyawaki S, et al. A markov decision analysis of allogeneic hematopoietic cell transplantation versus chemotherapy in patients with acute myeloid leukemia in first remission. Blood. 2011;117(7):2113-2120.
197. Greiner RA, Meier Y, Papadopoulos G, O'Sullivan AK, Imhof A. Cost-effectiveness of posaconazole compared with standard azole therapy for prevention of invasive fungal infections in patients at high risk in switzerland. Oncology. 2010;78(3-4):172-180.
198. Song KW, Lipton J. Is it appropriate to offer allogeneic hematopoietic stem cell transplantation to patients with primary refractory acute myeloid leukemia? Bone Marrow Transplant. 2005;36(3): 183-191.
199. McCabe C, Dixon S. Testing the validity of cost-effectiveness models. Pharmacoeconomics. 2000; 17(5):501-513.
200. Hammerschmidt T, Goertz A, Wagenpfeil S, Neiss A, Wutzler P, Banz K. Validation of health economic models: The example of EVITA. Value Health. 2003;6(5):551-559.
201. Smith ML, Hills RK, Grimwade D. Independent prognostic variables in acute myeloid leukaemia. Blood Rev. 2011;25(1):39-51.
202. Greenwood MJ, Seftel MD, Richardson C, et al. Leukocyte count as a predictor of death during remission induction in acute myeloid leukemia. Leuk Lymphoma. 2006;47(7):1245-1252.
203. Larson RA. Is secondary leukemia an independent poor prognostic factor in acute myeloid leukemia? Best Pract Res Clin Haematol. 2007;20(1):29-37.
204. Lowenberg B, Griffin JD, Tallman MS. Acute myeloid leukemia and acute promyelocytic leukemia. Hematology. 2003;2003(1):82-101.
205. Vellenga E, van Putten W, Ossenkoppele GJ, et al. Autologous peripheral blood stem cell transplantation for acute myeloid leukemia. Blood. 2011;118(23):6037-6042.
206. Craddock C, Tauro S, Moss P, Grimwade D. Biology and management of relapsed acute myeloid leukaemia. Br J Haematol. 2005;129(1):18-34.
207. Breems DA, Van Putten WL, Huijgens PC, et al. Prognostic index for adult patients with acute myeloid leukemia in first relapse. J Clin Oncol. 2005;23(9):1969-1978.

208. Heeg BM, Damen J, Buskens E, Caleo S, de Charro F, van Hout BA. Modelling approaches: The case of schizophrenia. Pharmacoeconomics. 2008;26(8):633-648.
209. Kern W, Haferlach T, Schoch C, et al. Early blast clearance by remission induction therapy is a major independent prognostic factor for both achievement of complete remission and long-term outcome in acute myeloid leukemia: Data from the german AML cooperative group (AMLCG) 1992 trial. Blood. 2003;101(1):64-70.
210. Goldhaber-Fiebert JD, Stout NK, Goldie SJ. Empirically evaluating decision-analytic models. Value Health. 2010;13(5):667-674.
211. Integraal Kankercentra Nederland. Cijfers over kanker. Available from http://www.cijfersoverkanker.nl/. [Accessed 08/16, 2012].
212. Burnett AK, Hills RK, Milligan D, et al. Identification of patients with acute myeloblastic leukemia who benefit from the addition of gemtuzumab ozogamicin: Results of the MRC AML15 trial. Journal of Clinical Oncology. 2011;29(4):369-377.
213. Lee J, Joo Y, Kim H, et al. A randomized trial comparing standard versus high-dose daunorubicin induction in patients with acute myeloid leukemia. Blood. 2011;118(14):3832-3841.
214. Mandelli F, Vignetti M, Suciu S, et al. Daunorubicin versus mitoxantrone versus idarubicin as induction and consolidation chemotherapy for adults with acute myeloid leukemia: The EORTC and GIMEMA groups study AML-10. Journal of Clinical Oncology. 2009;27(32):5397-5403.
215. Ohtake S, Miyawaki S, Fujita H, et al. Randomized study of induction therapy comparing standard-dose idarubicin with high-dose daunorubicin in adult patients with previously untreated acute myeloid leukemia: The JALSG AML201 study. Blood. 2011;117(8):2358-2365.
216. Wheatley K, Goldstone AH, Littlewood T, Hunter A, Burnett AK. Randomized placebo-controlled trial of granulocyte colony stimulating factor (G-CSF) as supportive care after induction chemotherapy in adult patients with acute myeloid leukaemia: A study of the united kingdom medical research council adult leukaemia working party. Br J Haematol. 2009;146(1):54-63.
217. Kolitz JE, George SL, Marcucci G, et al. P-glycoprotein inhibition using valspodar (PSC-833) does not improve outcomes for patients younger than age 60 years with newly diagnosed acute myeloid leukemia: Cancer and leukemia group B study 19808. Blood. 2010;116(9):1413-1421.
218. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices task Force–7. Medical Decision Making. September–October 2012;32(5):733-743.
219. Choudhury AD, Eeles R, Freedland SJ, et al. The role of genetic markers in the management of prostate cancer. Eur Urol. 2012.
220. Voora D, Ginsburg GS. Clinical application of cardiovascular pharmacogenetics. J Am Coll Cardiol. 2012;60(1):9-20.
221. Karnon J. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. Health Econ. 2003;12(10):837-848.
222. Simpson KN, Strassburger A, Jones WJ, Dietz B, Rajagopalan R. Comparison of markov model and discrete-event simulation techniques for HIV. Pharmacoeconomics. 2009;27(2):159-165.
223. Li Z, Herold T, He C, et al. Identification of a 24-gene prognostic signature that improves the european LeukemiaNet risk classification of acute myeloid leukemia: An international collaborative study. Journal of Clinical Oncology. 2013.
224. Ludwig H, Durie BGM, McCarthy P, et al. IMWG consensus on maintenance therapy in multiple myeloma. Blood. 2012;119(13):3003-3015.
225. Roden DM, Altman RB, Benowitz NL, et al. Pharmacogenomics: Challenges and opportunities. Annals of Internal Medicine. 2006;145(10):749-757.

226. Weinshilboum R. Inheritance and drug response. N Engl J Med. 2003;348(6):529-537.
227. Tuckson RV, Newcomer L,De Sa JM. Accessing genomic medicine: Affordability, diffusion, and disparities. JAMA. 2013;309(14):1469-1470.
228. Garber AM, Tunis SR. Does comparative-effectiveness research threaten personalized medicine? N Engl J Med. 2009;360(19):1925-1927.
229. Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. Assessing personalized medicines in australia: A national framework for reviewing codependent technologies. Medical Decision Making. 2013;33(3):333-342.
230. Wong WB, Carlson JJ, Thariani R, Veenstra DL. Cost effectiveness of pharmacogenomics: A critical and systematic review. Pharmacoeconomics. 2010;28(11):1001-1013.
231. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. Journal of the National Cancer Institute. 2009;101(21):1446-1452.
232. Parkinson DR, McCormack RT, Keating SM, et al. Evidence of clinical utility: An unmet need in molecular diagnostics for patients with cancer. Clinical Cancer Research. 2014;20(6):1428-1444.
233. Hartung DM, Touchette D. Overview of clinical research design. American Journal of Health-System Pharmacy. 2009;66(4):398-408.
234. Parkinson B, Pearson S, Viney R. Economic evaluations of trastuzumab in HER2-positive metastatic breast cancer: A systematic review and critique. The European Journal of Health Economics. 2014; 15(1):93-112.
235. Phillips KA. CLosing the evidence gap in the use of emerging testing technologies in clinical practice. JAMA. 2008;300(21):2542-2544.
236. Parthan A, Leahy KJ, O'Sullivan AK, et al. Cost effectiveness of targeted high-dose atorvastatin therapy following genotype testing in patients with acute coronary syndrome. Pharmacoeconomics. 2013;31(6):519-531.
237. Retèl VP, Joore MA, Knauer M, Linn SC, Hauptmann M, Harten WHv. Cost-effectiveness of the 70-gene signature versus st. gallen guidelines and adjuvant online for early breast cancer. Eur J Cancer. 2010;46(8):1382-1391.
238. Nelson RE, Stenehjem D, Akerley W. A comparison of individualized treatment guided by VeriStrat with standard of care treatment strategies in patients receiving second-line treatment for advanced non-small cell lung cancer: A cost-utility analysis. Lung Cancer. 2013;82(3):461-468.
239. Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS)—Explanation and elaboration: A report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. *Value Health*. 2013;16(2): 231-250.
240. Ramsey S, Willke R, Briggs A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: The ISPOR RCT-CEA task force report. Value Health. 2005;8(5):521-533.
241. Hall P, McCabe C. What evidence is there for the reimbursement of personalised medicine? Pharmacoeconomics. 2013;31(3):181-183.
242. Annemans L, Redekop K, Payne K. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(6 Suppl):S20-6.
243. Leunis A, Redekop WK, van Montfort KA, Lowenberg B, Uyl-de Groot CA. The development and validation of a decision-analytic model representing the full disease course of acute myeloid leukemia. Pharmacoeconomics. 2013;31(7):605-621.
244. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998;17(24):2815-2834.

245. Atherly AJ, Camidge DR. The cost-effectiveness of screening lung cancer patients for targeted drug sensitivity markers. Br J Cancer. 2012;106(6):1100-1106.
246. Djalalov S, Beca J, Hoch JS, et al. Cost effectiveness of EML4-ALK fusion testing and first-line crizotinib treatment for patients with advanced ALK-positive non-small-cell lung cancer. J Clin Oncol. 2014;32(10):1012-1019.
247. Frank M, Mittendorf T. Influence of pharmacogenomic profiling prior to pharmaceutical treatment in metastatic colorectal cancer on cost effectiveness : A systematic review. Pharmacoeconomics. 2013;31(3):215-228.
248. Greeley SAW, John PM, Winn AN, et al. The cost-effectiveness of personalized genetic medicine: The case of genetic testing in neonatal diabetes. Diabetes Care. 2011;34(3):622-627.
249. Lidgren M, Jönsson B, Rehnberg C, et al. Cost-effectiveness of HER2 testing and 1-year adjuvant trastuzumab therapy for early breast cancer. *Annals of Oncology*. 2008;19(3):487-495.
250. Jönsson B. Bringing in health technology assessment and cost-effectiveness considerations at an early stage of drug development. Molecular Oncology. (0).
251. Mrózek K, Heerema NA, Bloomfield CD. Cytogenetics in acute leukemia. Blood Rev. 2004;18(2): 115-136.
252. Ferrara F, Schiffer CA. Acute myeloid leukaemia in adults. The Lancet. 2013;381(9865):484-495.
253. Dufour A, Schneider F, Metzeler KH, et al. Acute myeloid leukemia with biallelic CEBPA gene mutations and normal karyotype represents a distinct genetic entity associated with a favorable clinical outcome. Journal of Clinical Oncology. 2010;28(4):570-577.
254. Green CL, Koo KK, Hills RK, Burnett AK, Linch DC, Gale RE. Prognostic significance of CEBPA mutations in a large cohort of younger adult patients with acute myeloid leukemia: Impact of double CEBPA mutations and the interaction with FLT3 and NPM1 mutations. J Clin Oncol. 2010;28(16): 2739-2747.
255. Boissel N, Renneville A, Biggio V, et al. Prevalence, clinical profile, and prognosis of NPM mutations in AML with normal karyotype. Blood. 2005;106(10):3618-3620.
256. Döhner K, Schlenk RF, Habdank M, et al. Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: Interaction with other gene mutations. Blood. 2005;106(12):3740-3746.
257. Thiede C, Koch S, Creutzig E, et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). Blood. 2006;107(10):4011-4020.
258. Groschel S, Lugthart S, Schlenk RF, et al. High EVI1 expression predicts outcome in younger adult patients with acute myeloid leukemia and is associated with distinct cytogenetic abnormalities. J Clin Oncol. 2010;28(12):2101-2107.
259. Schlenk RF, Pasquini MC, Perez WS, et al. HLA-identical sibling allogeneic transplants versus chemotherapy in acute myelogenous leukemia with t(8;21) in first complete remission: Collaborative study between the german AML intergroup and CIBMTR. Biol Blood Marrow Transplant. 2008; 14(2):187-196.
260. Nathan PC, Sung L, Crump M, Beyene J. Consolidation therapy with autologous bone marrow transplantation in adults with acute myeloid leukemia: A meta-analysis. J Natl Cancer Inst. 2004; 96(1):38-45.
261. Breems DA, Boogaerts MA, Dekker AW, et al. Autologous bone marrow transplantation as consolidation therapy in the treatment of adult patients under 60 years with acute myeloid leukaemia in first complete remission: A prospective randomized dutch-belgian haemato-oncology cooperative group (HOVON) and swiss group for clinical cancer research (SAKK) trial. Br J Haematol. 2005;128(1):59-65.

262. Ringden O, Labopin M, Gorin NC, et al. Is there a graft-versus-leukaemia effect in the absence of graft-versus-host disease in patients undergoing bone marrow transplantation for acute leukaemia? Br J Haematol. 2000;111(4):1130-1137.
263. Suciu S, Mandelli F, de Witte T, et al. Allogeneic compared with autologous stem cell transplantation in the treatment of patients younger than 46 years with acute myeloid leukemia (AML) in first complete remission (CR1): An intention-to-treat analysis of the EORTC/GIMEMAAML-10 trial. Blood. 2003;102(4):1232-1240.
264. Brunet S, Esteve J, Berlanga J, et al. Treatment of primary acute myeloid leukemia: Results of a prospective multicenter trial including high-dose cytarabine or stem cell transplantation as postremission strategy. Haematologica. 2004;89(8):940-949.
265. Lazarus HM, Perez WS, Klein JP, et al. Autotransplantation versus HLA-matched unrelated donor transplantation for acute myeloid leukaemia: A retrospective analysis from the center for international blood and marrow transplant research. Br J Haematol. 2006;132(6):755-769.
266. Leunis A, Blommestein HM, Huijgens PC, Blijlevens NMA, Jongen-Lavrencic M, Uyl-de Groot CA. The costs of initial treatment for patients with acute myeloid leukemia in the netherlands. Leuk Res. 2013;37(3):245-250.
267. van Agthoven M, Groot MT, Verdonck LF, et al. Cost analysis of HLA-identical sibling and voluntary unrelated allogeneic bone marrow and peripheral blood stem cell transplantation in adults with acute myelocytic leukaemia or acute lymphoblastic leukaemia. Bone Marrow Transplant. 2002; 30(4):243-251.
268. Schneider F, Hoster E, Unterhalt M, et al. NPM1 but not FLT3-ITD mutations predict early blast cell clearance and CR rate in patients with normal karyotype AML (NK-AML) or high-risk myelodysplastic syndrome (MDS). Blood. 2009;113(21):5250-5253.
269. Schlenk RF, Döhner H. Genomic applications in the clinic: Use in treatment paradigm of acute myeloid leukemia. ASH Education Program Book. 2013;2013(1):324-330.
270. Kekre N, Koreth J. Novel strategies to prevent relapse after allogeneic haematopoietic stem cell transplantation for acute myeloid leukaemia and myelodysplastic syndromes. Curr Opin Hematol. 2015;22(2):116-122.
271. Brazier J, Connell J, Papaioannou D, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess. 2014;18(34):vii-viii, xiii-xxv, 1-188.
272. Courneya KS, McKenzie DC, Mackey JR, et al. Moderators of the effects of exercise training in breast cancer patients receiving chemotherapy. Cancer. 2008;112(8):1845-1853.
273. Courneya KS, Sellar CM, Stevinson C, et al. Moderator effects in a randomized controlled trial of exercise training in lymphoma patients. Cancer Epidemiology Biomarkers & Prevention. 2009; 18(10):2600-2607.
274. Bergenthal N, Will A, Streckmann F, et al. Aerobic physical exercise for adult patients with haematological malignancies. Cochrane Database Syst Rev. 2014;11:CD009075.
275. Mishra SI, Scherer RW, Geigle PM, et al. Exercise interventions on health-related quality of life for cancer survivors. Cochrane Database Syst Rev. 2012;8:CD007566.
276. Nederlandse Zorgautoriteit. DBC zorgproducten tariefapplicatie. Available from http://dbc-zorgproducten-tarieven.nza.nl/nzaZpTarief/Welkom.aspx. Updated 2015. [Accessed February/10, 2015].
277. Gaultney JG, Franken MG, Tan SS, et al. Real-world health care costs of relapsed/refractory multiple myeloma during the era of novel cancer agents. J Clin Pharm Ther. 2013;38(1):41-47.

278. College voor Zorgverzekeringen. Richtlijnen voor farmaco-economisch onderzoek, geactualiseerde versie. 2006. ; No. 25001605Available at http://www.zorginstituutnederland.nl/binaries/content/documents/zinl-www/documenten/rubrieken/pakket/pakketbeheer/0604-richtlijnen-voor-farmaco-economisch-onderzoek/Richtlijnen+voor+farmaco-economisch+onderzoek.pdf.
279. Uyl-de Groot CA. 'Dure' diagnostiek en kankergeneesmiddelen: De andere kant van de ongelijkheid II. 2011. Available at http://www.bmg.eur.nl/fileadmin/ASSETS/bmg/Onderzoek/Oraties/Uyl/200511_oratie_uyl.pdf.
280. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: Design issues. Journal of the National Cancer Institute. 2010;102(3):152-160.
281. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of Clinical Oncology. 2005;23(9):2020-2027.
282. Wason J, Marshall A, Dunn J, Stein RC, Stallard N. Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. Br J Cancer. 2014;110(8):1950-1957.
283. Hills RK, Burnett AK. Applicability of a "pick a winner" trial design to acute myeloid leukemia. Blood. 2011;118(9):2389-2394.
284. Juliusson G, Lazarevic V, Hörstedt A, Hagberg O, Höglund M. Acute myeloid leukemia in the real world: Why population-based registries are needed. Blood. 2012;119(17):3890-3899.
285. Blommestein HM, Franken MG, Uyl-de Groot CA. A practical guide for using registry data to inform decisions about the cost effectiveness of new cancer drugs: Lessons learned from the PHAROS registry. Pharmacoeconomics. 2015.
286. Hunink M, Glasziou P, Siegel J, et al. Interpreting diagnostic information. In: Decison making in health and medicine. Cambridge: University Press; 2001:128-156.
287. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. American Journal of Epidemiology. 2012; 176(6):473-481.
288. Duncan RE, Gillam L, Savulescu J, Williamson R, Rogers JG, Delatycki MB. "You're one of us now": Young people describe their experiences of predictive genetic testing for huntington disease (HD) and familial adenomatous polyposis (FAP). American Journal of Medical Genetics Part C: Seminars in Medical Genetics. 2008;148C(1):47-55.
289. Parens E, Appelbaum P, Chung W. Incidental findings in the era of whole genome sequencing? Hastings Cent Rep. 2013;43(4):16-19.
290. Kohane IS, Masys DR,Altman RB. The incidentalome: A threat to genomic medicine. JAMA. 2006; 296(2):212-215.
291. IJzerman MJ, Steuten LM. Early assessment of medical technologies to inform product development and market access: A review of methods and applications. Appl Health Econ Health Policy. 2011;9(5):331-347.