TRANSLATING EVIDENCE INTO VALUE FOR STAKEHOLDERS

A STRUCTURED APPROACH USING
PROBABILISTIC MULTI-CRITERIA DECISION ANALYSIS

Henk Broekhuizen

The colouring pencils on the cover are an analogy for the methods developed in this thesis. Clinical evidence describes the effects of interventions, but knowledge about the subjective/colored patient experience is needed to interpret this objective data. There are multiple pencils on the cover because the thesis explores uncertainty and variation that affect the patient experience.

Translating evidence into value for stakeholders:

A Structured Approach using Probabilistic Multi-criteria Decision Analysis

ISBN: 978-90-365-4354-5

DOI: 10.3990/1.9789036543545

This thesis is part of the Health Sciences Series of the department for Health Technology and Services Research at the University of Twente (Enschede, the Netherlands), HSS 17-15, ISSN 1878-4968.

Printed by Gildeprint Enschede on FSC certified paper

© Copyright 2017: Henk Broekhuizen, Enschede, the Netherlands

All rights reserved. No part of this publication may be reproduced without permission of the copyright holder.

TRANSLATING EVIDENCE INTO VALUE FOR STAKEHOLDERS

A STRUCTURED APPROACH USING
PROBABILISTIC MULTI-CRITERIA DECISION ANALYSIS

Dissertation

to obtain the degree of doctor at the University of Twente
on the authority of the rector magnificus,
Prof. dr. T.T.M. Palstra
in accordance with the decision of the graduation committee,
to be defended in public
on Thursday 15th, June 2017, at 14:45 hours

by

Hindrik Broekhuizen

born on 3rd October 1987
in Hengelo (OV), The Netherlands

This dissertation has been approved by:

Prof. dr. M.J. IJzerman (supervisor)

Dr. C.G.M. Groothuis-Oudshoorn (co-supervisor)

Graduation Committee

Chairman/secretary

Prof. dr. Th.A.J. Toonen University of Twente

Supervisor

Prof. dr. M.J. IJzerman University of Twente

Co-supervisor

Dr. C.G.M. Groothuis-Oudshoorn University of Twente

Referee

Dr. D. Postmus University Medical Center Groningen

Members

Prof. dr. R.J. Boucherie University of Twente

Prof. dr. S. Siesling University of Twente

Prof. dr. A.C. Muhlbacher Hochschule Neubrandenburg

Prof. dr. C.D. Dirksen Maastricht University Medical Center

Paranymphs

Björn Postema

Tjark van de Merwe

Table of contents

Chapter 1: <i>General introduction</i>	1
Chapter 2: <i>A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions</i>	9
Chapter 3: <i>Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis</i>	31
Chapter 4: <i>Weighing clinical evidence using patient preferences: an application of probabilistic multi-criteria decision analysis</i>	49
Chapter 5: <i>A health economic perspective on the future of lung cancer screening</i>	71
Chapter 6: <i>Public preferences for lung cancer screening policies</i>	83
Chapter 7: <i>Modeling multivariate preference distributions using copula functions</i>	103
Chapter 8: <i>Assessing the value of lung cancer screening programs under uncertainty in a heterogeneous population</i>	127
Chapter 9: <i>General Discussion</i>	149
Summary (English and Dutch)	159
CV auteur en dankwoord	163

Chapter 1

General introduction

Background and aims

Setting

For several decades, evidence-based medicine (EBM) has been advocated to support clinicians in taking care for patients [1]. The central tenet underlying EBM is that the available clinical evidence should be used in a conscientious, explicit, and judicious manner. While EBM is still the leading paradigm for clinical decisions, new initiatives and methods are required to deal with the complexity of healthcare delivery and the requirements of stakeholders. Health technology assessment (HTA) and comparative effectiveness research (CER) are built on the foundations of EBM but aim to improve patient outcomes and processes of care delivery across the whole spectrum of healthcare instead of only at a clinical level [2]. In HTA and CER, evidence¹ from different sources is combined to help a variety of decision makers (patients, clinicians, payers, and policy makers) make informed decisions based on the best available evidence. While this applies to clinical decisions, it also is relevant for drug and medical device regulation (i.e. market access and reimbursement), for public health campaigns, and for prevention programs.

In recent years, healthcare policy decisions have been criticized for a lack of transparency regarding how the decisions were reached given the available evidence [3]. The distinction between confirmed clinical evidence and expert judgment is also not always clear and this is especially a relevant concern when large amounts of clinical evidence are being evaluated [4]. Given the criticisms, it is not surprising that decision makers are increasingly looking at quantitative decision models to guide evidence-synthesis and decision-making processes [5].

The role of evidence in healthcare policy decision making

Because clinical evidence is seldom a single piece of information, it is more useful to speak about an *evidence base* consisting of multiple pieces of evidence. Examples of such pieces of evidence are empirical evidence from randomized trials, observational trials using multiple clinical outcome measures, patient registries, results from consultation processes, input from clinical experts, political information regarding opportunity and crises, and

¹ In this thesis we define clinical evidence to be “clinically relevant research, often from the basic sciences of medicine, but especially from patient centered clinical research into the accuracy and precision of diagnostic tests (including the clinical examination), the power of prognostic markers, and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens.”[17]

economic information [6]. Normative aspects of how evidence is used in policy-making have undergone considerable debate [7]. However, it is generally agreed that evidence is used by regulators to update their prior beliefs about the outcomes of the decision at hand [8]. How the various pieces of evidence are combined is called *evidence synthesis*, of which a meta-analysis as part of systematic reviews is a well-known example. Evidence synthesis yields objective information about clinical outcomes, but that alone is not enough to actually make the decision. Inevitably, value judgments must be made about what these outcomes mean for those who will experience them [9]. Consider for example the trade-off that has to be made between the beneficial and adverse outcomes of a drug or between various types of adverse outcomes. What trade-offs are acceptable for someone is a personal and highly subjective value judgment about the clinical evidence base. There is no scientific study that can prove that it is ‘correct’, for example, to be risk averse or risk neutral.

Stated-preference methods to gather stakeholder preferences

The main challenge in healthcare policy decision making, therefore, is to make value judgments over multiple sources of evidence about the relative benefits and harms of interventions while taking into account the perspective of the most important stakeholders, i.e. patients or participants in public health programs². Over the years, several frameworks have been proposed and used to estimate the value of interventions for these stakeholders [10]. One example of a patient-based value-framework is cost-effectiveness analyses, where quality-adjusted life years represent the value of an intervention. However, the QALY may not capture all elements that people ultimately value such as equity considerations or the process of care delivery [11]. By expanding value frameworks to incorporate these elements, a more holistic view of what stakeholders value can be obtained. Consequently, this enables policy makers to make decisions that are in accordance with stakeholder needs. Although individual patient representatives have been involved in for example regulatory committees, stated preference methods are increasingly seen as a more representative approach for gaining insight into what stakeholders value [12]. Stated preference methods are structured and transparent methods that can inform policy makers on what attributes of interventions are valued by stakeholders and what tradeoffs stakeholders would be willing to make between attributes [13]. The main advantage of stated preference methods is that they are survey-based, which makes it possible to measure the preferences of a large group of stakeholders. Preferential evidence obtained with stated preference methods may be used by policy makers to update their

² From now on we will use the term ‘stakeholder’ to denote these two groups.

prior values, much in the same way as clinical evidence can be used to update their prior beliefs about the outcomes.

Multi-criteria decision analysis to translate evidence into value

In this thesis we pose that multi-criteria decision analysis (MCDA) is useful for assessing the value of interventions by different stakeholders. How MCDA can be used becomes clear if the four main steps taken during an MCDA are considered: structuring, scoring, weighting, and aggregating [14]. In the first ‘structuring’ step, the decision context and the set of attributes on which to evaluate the intervention(s) are identified and structured. Attributes are mostly identified through literature and expert consultations, and usually include clinical attributes (such as symptom reduction or side effects) as well as attributes relating to the process of healthcare delivery (such as waiting time or mode of administration). In this way, a framework of what constitutes value for stakeholders is established in a structured and explicit manner. In the second step of MCDA, clinical evidence or expert opinion is used to assign numeric scores to the performance of each intervention on all included attributes. This means it is explicitly recorded what sources of evidence were used and how these sources are thought to inform the performance of interventions. After scoring, each attribute is assigned an importance weight that indicates its importance relative to the other attributes. The relative importance of attributes is subjective and could be informed by including results from stated preference research. By doing so, the views of large groups of stakeholders are included into the value assessment in a clear and structured manner. In the final step of MCDA, scores and weights are aggregated into an overall value for each choice alternative. When in the scoring and weighting steps clinical evidence was used to inform the scores and when stated preference research was used to inform the weights, the obtained overall values indicate the evidence-based value of interventions according to stakeholders.

This thesis

Rationale: MCDA for integrating evidence

and value assessment under uncertainty

While the integration of stakeholder views and clinical evidence in one value framework is the main objective of the present thesis, the analytic approach to combine clinical evidence and stated preference research is particularly challenging because of the impact of uncertainty in the evidence and preferences on the outcomes. Here we distinguish stochastic uncertainty, heterogeneity, parameter uncertainty, and structural uncertainty

[15]. Stochastic uncertainty is random variability between otherwise identical people. Stochastic uncertainty in (preferences for) clinical outcomes should be considered to enable policy makers to understand the proportion of patients for whom the benefits of an intervention may outweigh the risks [16]. This is even more poignant if different treatment outcomes can be attributed to observable patient characteristics like age or genetic profile, i.e. if there is heterogeneity. Parameter uncertainty in preferences or clinical evidence may be reduced by more research but the question is if more research and thus less parameter uncertainty would actually lead to different decisions. Finally, structural uncertainty can exist, for example, about what attributes of interventions to include in the decision. Not taking into account these types of uncertainty may imply decisions are taken that are not in line with what is valued by stakeholders.

Aim and outline

The aim of this thesis is to investigate how the various types of uncertainty impact the outcomes of a stakeholder-based value framework for comparing medical interventions. Our MCDA-based approach will be developed over three sections.

First, we will investigate how previous studies have considered uncertainty in the context of MCDA. For this purpose, **Chapter 2** will introduce an MCDA-specific theoretical framework of uncertainty and review the scientific literature to identify approaches for dealing with the various types of uncertainty. Of the five identified approaches, we consider the probabilistic approach the most suitable for our aims mainly because it allows for the simultaneous consideration of multiple sources of uncertainty.

In the second section, we will further develop the probabilistic MCDA approach for building stakeholder-based value frameworks. In **Chapter 3**, parameter uncertainty in the clinical performance of antidepressants is weighted using a simulated patient preference dataset. Monte Carlo simulations are introduced as a method for combining the uncertainty from multiple sources. **Chapter 4** builds on this by extending the MCDA model of chapter 3 to also include stochastic uncertainty in preferences for the case of antiretroviral treatments. Both chapters will investigate the relative impact of the types of uncertainty by performing scenario analyses.

In the third section a stakeholder-based value framework is developed for the case of lung cancer screening. This application area is introduced in **Chapter 5** where we discuss the notion that preference-related barriers to entry into a voluntary screening program for lung cancer are important but under-researched. **Chapter 6** investigates preferences for

attributes of lung cancer screening programs according to a large public sample from the Dutch population. Heterogeneity in preferences for these attributes is explored by identifying five preference subgroups using hierarchical cluster analysis. How correlations in preference data between preferences for different attributes can be taken into account in a probabilistic MCDA is investigated in **Chapter 7**, where we introduce copulas for modeling multivariate preference data and compare the copula approach to conventional methods. In **Chapter 8** we apply the probabilistic MCDA model developed in the second section to the case of lung cancer screening. For this, we gather clinical evidence about screening programs and combine this evidence with the preference data gathered in chapter 6 to obtain value estimates for three different screening programs. In our analyses we model parameter uncertainty in the estimates and take into account heterogeneity in preferences by estimating the value of the screening programs according to various subgroups in the public sample.

Finally, **Chapter 9** provides a discussion of the work performed in this thesis, relates it to the broader literature, and suggests avenues for further research.

References

- [1] Sur R, Dahm P. The history of evidence-based medicine. *Indian J Urol* 2011;27:487–9.
- [2] Sox H, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med* 2009;151:203–5.
- [3] Yuan Z, Levitan B, Berlin JA. Benefit-risk assessment: to quantify or not to quantify, that is the question. *Pharmacoepidemiol Drug Saf* 2011;20:653–6.
- [4] Phillips LD, Fasolo B, Zafriopoulos N, Beyer A. Is quantitative benefit–risk modelling of drugs desirable or possible? *Drug Discov Today Technol* 2011;8:e3–10.
- [5] Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. *Value Heal* 2010;13:657–66.
- [6] Bowen S, Zwi AB. Pathways to “evidence-informed” policy and practice: A framework for action. *PLoS Med* 2005;2:0600–5.
- [7] Hunter DJ. Relationship between evidence and policy: a case of evidence-based policy or policy-based evidence? *Public Health* 2009;123:583–6.
- [8] Cookson R. Evidence-based policy making in health care: what it is and what it isn’t. *J Health Serv Res Policy* 2005;10:118–21.
- [9] Eddy D. Anatomy of a decision. *JAMA* 1990;263:441–3.
- [10] Doshi J, Willke R. Advancing High-Quality Value Assessments of Health Care Interventions. *Value Heal* 2017;20:181–4.
- [11] Weernink MGM, Janus SIM, van Til J a., Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceut Med* 2014.
- [12] van Til JA, IJzerman MJ. Why Should Regulators Consider Using Patient Preferences in Benefit-risk Assessment? *Pharmacoeconomics* 2013;10–3.

CHAPTER 1

- [13] Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl Health Econ Health Policy* 2013;11:319–29.
- [14] Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making-An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Heal* 2015;19:1–13.
- [15] Briggs AH, Weinstein MC, Fenwick E a L, Karnon J, Sculpher MJ, Paltiel a D. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Heal* 2012;15:835–42.
- [16] Ho MP, Gonzalez JM, Lerner HP, Neuland CY, Whang JM, McMurry-Heath M, et al. Incorporating patient-preference evidence into regulatory decision making. *Surg Endosc* 2015.
- [17] Sacket D. Evidence-based medicine. *Semin Perinatol* 1997;21:3–5.

Chapter 2

A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions

Henk Broekhuizen

Catharina G.M. Groothuis-Oudshoorn

Janine A. van Til

J. Marjan Hummel

Maarten J. IJzerman

Published as: Broekhuizen H, Groothuis-Oudshoorn CGM, van Til JA, Hummel JM, IJzerman MJ. A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics* 2015;33:445–55.

Abstract

Multi-criteria decision analysis (MCDA) is increasingly used to support decisions in health care involving multiple and conflicting criteria. Although uncertainty is usually carefully addressed in health economics evaluations, whether and how the different sources of uncertainty are dealt with and with what methods in MCDA is less known. The objective of this study is to review how uncertainty can be explicitly taken into account in MCDA and to discuss which approach may be appropriate for healthcare decision makers. A literature review was conducted in the Scopus and Pubmed databases. Two reviewers independently categorized studies according to research areas, the type of MCDA used, and the approach used to quantify uncertainty. Selected full text articles were read for methodological details. The search strategy identified 569 studies. The five approaches most identified were fuzzy set theory (45% of studies), probabilistic sensitivity analysis (15%), deterministic sensitivity analysis (31%), Bayesian framework (6%) and grey theory (3%). A large number of papers considered the Analytic Hierarchy Process in combination with fuzzy set theory (31%). Only 3% of studies were published in healthcare-related journals. In conclusion, our review identified five different approaches to take uncertainty into account in MCDA. The deterministic approach is most likely sufficient for most health care policy decisions because of its low complexity and straightforward implementation. However, more complex approaches may be needed when multiple sources of uncertainty must be considered simultaneously.

Introduction

Over the last decade, researchers in outcomes research have increasingly suggested multi-criteria decision analysis (MCDA) as an approach to support healthcare decisions [1]. MCDA is an extension of decision theory that supports decision makers (policy makers, regulators, managers, etc.) who have multiple (possibly conflicting) objectives by decomposing the decision objectives into criteria [2]. These criteria are given a numeric importance weight and decision alternatives such as drugs or treatments are scored on each of the criteria. The criteria weights and performances scores are then aggregated into an overall score which is used to rank the alternative treatments. For a more detailed overview of the subsequent steps in MCDA, the reader is referred to Belton and Stewart [2] and Hummel *et al.* [3]. MCDA is considered to be a transparent and flexible approach [4–7]. MCDA has been used to support a wide range of decisions, such as in portfolio optimization, benefit-risk assessment, health technology assessment and shared decision making [8–13]. Although the objectives differ, these decisions share three characteristics. First, they are characterized by possibly conflicting decision criteria where trade-offs between criteria influence the decision. Second, the criteria to operationalize can be qualitative, quantitative, or a combination of both. And finally, these decisions and underlying criteria weights and performance scores are characterized by uncertainty.

In principal, several sources of uncertainty can be distinguished and have been clearly described by different authors [14,15]. In their comprehensive taskforce report, Briggs *et al.* define four types of uncertainty: stochastic uncertainty, parameter uncertainty, heterogeneity, and structural uncertainty [15]. Although their report discusses uncertainty in decision analytic models in general, this classification is almost identical for MCDA. However, in MCDA the four types of uncertainty are relevant to consider for both the weighting of criteria and the scoring of alternatives. Criteria weights are always elicited from decision makers, and stochastic uncertainty in weighting is therefore random variability in weights as assigned by otherwise identical persons. Parameter uncertainty refers to the estimation error of an estimated quantity, for example the mean weight given by a group of decision makers to a criterion. Heterogeneity is explainable variation in weights, for example due to a person's background characteristics. Finally, structural uncertainty occurs when decision makers are unsure if all relevant decision criteria are included and how these criteria are structured [14].

Type of uncertainty	Definition Briggs <i>et al.</i> [15]		MCDA-specific definition
	<i>Stochastic uncertainty</i>	Random variability in outcomes between identical patients	Random variability in criteria weights or performance scores as assigned by identical persons
	<i>Parameter uncertainty</i>	The uncertainty in estimation of the parameter of interest	The uncertainty in estimation of the parameter (criterion weight or performance score) of interest
	<i>Heterogeneity</i>	The variability between patients that can be attributed to characteristics of those patients	Variability in criteria weights or performance scores that can be attributed to a person's characteristics
	<i>Structural uncertainty</i>	The assumptions inherent in the decision model	Uncertainty about if all relevant criteria are included, if they are properly structured and which transformations are used

Table 1: overview of types and sources of uncertainty in the context of MCDA-supported decision making. On the left are the types of uncertainty, as introduced by Briggs *et al.* [15]. In MCDA, uncertainty is related to both the determination of criteria weights and performance scores. Criteria weights are always elicited from stakeholders or decision makers while performance scores can either be elicited from stakeholders or derived from other data sources such as registries and clinical trials.

Like criteria weights, the performance scores of alternatives can be obtained through elicitation. Alternatively, performance scores can be obtained from other data sources such as registries or clinical trials. If performance scores are elicited from decision makers, generally the same sources of uncertainty apply as in the weighting step. If data is obtained from other data sources (such as an odds ratio comparing two drugs derived from a clinical trial), stochastic uncertainty, parameter uncertainty, and heterogeneity stems from variation or uncertainty in the source data. Structural uncertainty, however, is relevant to consider in these instances as it refers to how the outcomes are measured and how the data is transformed to a performance score in the MCDA. Often, performances scores are assigned based on a structured appraisal of available evidence. In that case, both elicitation-specific and data-specific uncertainties are relevant, of which the mix depends on the amount of available evidence. An overview of the types of uncertainty and their source in MCDA is presented in Table 1.

A recent systematic review by Marsh *et al* identified 41 applications of MCDA in healthcare and found that decision makers are positive about the possibilities of MCDA but that guidance on its application is lacking [16]. Twenty-two studies considered uncertainty, predominantly with deterministic sensitivity analysis. Previous studies outside the area of healthcare reviewed approaches to take into account uncertainty in MCDA-supported decisions. Durbach and Stewart reviewed different approaches to take into account

uncertainty in the scoring of alternatives [17]. They identified: probability-based approaches, fuzzy numbers, risk-based approaches and scenario analysis. Finally, a review by Kangas and Kangas in the field of forestry identified the frequentist, Bayesian, evidential reasoning, fuzzy sets, probabilistic, and possibility theory approaches [18].

Although the use of MCDA is emerging and uncertainty clearly is a relevant issue in MCDA models, there currently is no guidance on how uncertainty should be taken into account. To account for uncertainty in MCDA, three separate steps are proposed. First, the sources of uncertainty need to be identified, followed by (2) an assessment of the magnitude of the uncertainty and finally (3) by an evaluation whether the uncertainty would eventually lead to a different decision.

The objective of the present study is threefold. First, the study aims to identify common approaches to account for uncertainty. The second objective is to classify the identified approaches according to their mathematical approach and according to how the estimates for uncertainty are derived. Finally, the approaches will be compared, and their applicability for healthcare decisions will be discussed. In this discussion the focus will be on approaches that deal with elicitation-related uncertainty.

Methods

Identification of studies

A literature search in the SCOPUS and Pubmed databases was performed for the period between 1960 and 2013 using the following search terms: *(MCDA OR multi criteria decision analysis) AND (methodological OR parameter OR structural OR stochastic OR subjective OR *) uncertainty, multi criteria decision analysis AND (sensitivity OR robustness OR scenario) analysis, and uncertainty AND X and sensitivity analysis AND X*. In these strings, an asterisk represents a wildcard that can be matched by any word, and X was replaced with each of the individual MCDA method names, written both in full as well as in abbreviated form (see also Table 3). In addition to the database search, we also searched reference lists. Non-English studies, studies that did not apply or discuss MCDA and studies with an application of MCDA where uncertainty was not taken into account were excluded.

Classification

Following the identification, all included studies were classified by research area. This was done by coding the publications (journals and conference proceedings) in which the studies were published with their top-level All Science Journal Classification (ASJC). If publications were associated with multiple classifications, all were used for the coding. To examine applications in healthcare, sub-level ASJC codes related to healthcare were used (available from authors on request).

Secondly, the studies were classified by the MCDA method used. Only MCDA methods which were identified twice or more were put in separate categories, all methods which were used only once were put in the “other” category. MCDA methods were separated into: value-based, outranking, reference-based or other/hybrid methods. Value-based methods construct a single overall value for each decision alternative. Low scores on one criterion can be compensated by higher scores on another criterion. In outranking methods, low scores on one criterion may not be compensated by higher scores on another criterion. Furthermore, incomparability between the performance scores of alternatives is allowed. Reference-based methods calculate the similarity of alternatives to an ideal and anti-ideal alternative. The categorization was performed independently by three reviewers (HB, CG, MH) and disagreements were resolved through discussion. Full text articles were accessed when the used MCDA method could not be identified from the abstract alone or in case of continued disagreement between reviewers.

Third, two reviewers (HB, CG) independently classified the studies by their approach to take into account uncertainty. An initial list of approaches was defined based on the authors’ past experiences. On this initial list were the deterministic, probabilistic, Bayesian, and fuzzy set approaches. Newly identified approaches were added to the list. For every unique combination of an MCDA method and an uncertainty approach, the most recent full text article was read. If needed, references of these articles were also read to find methodological details or to identify textbooks.

Results from the literature review

Identification and classification of studies

A total of 569 studies were identified which were published between 1986 and 2013. The number of published studies increases sharply after the year 2000 (Figure 1). Top-level ASJC

research areas which accounted for more than 10% of the included studies are engineering (21%), computer science (17%) and environmental science (12%), as presented in Table 2. Only 3% of the included studies are published in a publication with an ASJC related to healthcare. Most studies (88%) use one MCDA method while 11% use two and 1% uses three. As seen in Table 3, the Analytic Hierarchy Process (AHP) was used most often (52%), followed by the Technique for Order of Preferences by Similarity to Ideal Solution (TOPSIS) (9%) and the Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) (7%). For 15% of studies the MCDA method used could not be identified from the abstract alone and therefore required full text reading.

Description of identified approaches for uncertainty analysis

Following the identification of studies, five distinctly different approaches were identified: deterministic sensitivity analysis, probabilistic sensitivity analysis, Bayesian frameworks, fuzzy set theory and grey theory. Fuzzy set theory was most commonly identified (45% of studies). The frequency with which the identified approaches were used in various research and application fields is presented in Table 2, and Table 3 presents how often the uncertainty approaches were combined with various MCDA methods. For a comprehensive demonstration of the different approaches, please be referred to the Excel file in the supplementary material found with the online version of the published paper based on this chapter.

Deterministic sensitivity analysis

Three types of deterministic sensitivity analyses were identified: simple sensitivity analysis, threshold analysis and analysis of extremes [19]. In simple sensitivity analysis, one model parameter (a criteria weight or a performance score) is varied at a time, and the impact of variation on the rank order of alternatives is observed. If the induced variation does not change the rank order of alternatives (i.e. the preference of one alternative over the other), the decision is considered robust [20,21]. Both threshold analysis and analysis of extremes are aimed at determining how much model parameters need to change before a different rank order of alternatives is obtained [22].

Probabilistic sensitivity analysis

Probabilistic sensitivity analysis requires decision makers to define probability distributions for model parameters. To assign probability distributions for performance scores, decision makers can refer to descriptive statistics or patient-level data from patient registries or clinical trials [23,24]. Methods to formally elicit probability distributions from (clinical) experts are available [25].

ASJC research area	Bayesian framework	DSA	Fuzzy set theory	Grey theory	PSA	n (%)
Agricultural and Biological Sciences	-	8	4	-	5	17 (2%)
Business, Management and Accounting	4	27	31	3	15	80 (8%)
Chemical Engineering	-	4	3	-	3	10 (1%)
Chemistry	-	1	4	-	4	9 (1%)
Computer Science	11	35	103	4	12	165 (17%)
Decision Sciences	8	33	30	2	27	100 (10%)
Earth and Planetary Sciences	3	3	5	-	2	13 (1%)
Economics, Econometrics and Finance	-	8	11	1	1	21 (2%)
Energy	-	20	5	1	4	30 (3%)
Engineering	9	56	121	8	17	211 (21%)
Environmental Science	3	43	42	1	26	115 (12%)
Materials Science	-	9	10	-	1	20 (2%)
Mathematics	8	26	43	2	22	101 (10%)
Medicine	1	10	8	-	4	23 (2%)
Physics and Astronomy	-	1	3	1	2	7 (1%)
Social Sciences	3	13	19	-	9	44 (4%)
Other	-	6	7	1	3	17 (2%)

Table 2: Research areas of publications in which identified studies were published, as coded with the all science journal classification (ASJC), and the division of identified approaches over the research areas. Note that a publication can be associated with more than one ASJC. DSA=deterministic sensitivity analysis. PSA=probabilistic sensitivity analysis

For model parameters for which there is little or no evidence, non-informative distributions such as a uniform distribution can be used [26].

After propagating the uncertainty through the MCDA model with Monte Carlo simulations, probability distributions are obtained for each alternative's overall score, and probabilistic statements such as the probability of a particular rank order of alternatives can be made [26–28].

Bayesian framework: Bayesian networks and Dempster-Shafer theory

Both Bayesian networks and Dempster-Shafer theory use the Bayesian framework to estimate the impact of uncertainty on the outcome of an MCDA-supported decision. Bayesian networks allow decision makers to explicitly model the interdependency of decision-related elements (for example, patient characteristics that may impact treatment performance) as a directed acyclic graph. Associated with each node in the graph are edges (arrows between nodes) that show the conditional relationships between the node and its parents. If there is no information about the conditional probabilities, prior distributions can be assumed. With more evidence, these prior probabilities can be updated using Bayes' theorem [29,30].

Figure 1: Distribution of identified studies by their year of publication and divided into the approach in which they were categorized.

Dempster-Shafer theory is an evidential reasoning-based method which is an extension of the Bayesian framework [31–33]. The five basic elements of Dempster-Shafer theory are: the frame of discernment, probability mass assignments, belief functions, plausibility functions and Dempster's rule of combination. The frame of discernment is a set of hypotheses from which one hypothesis with the most evidential support has to be chosen. In the context of MCDA, choice alternatives can be considered as the hypotheses; and the aim becomes to select the choice alternative for which the most evidential support for being the best choice (either for the whole decision or for a specific criterion) exists [34]. The first step in such an elicitation process consists of decision makers assigning a probability mass to (sets of) hypotheses in the frame of discernment, for example by

indicating that there is evidential support for treatment *A* and *B* being the best performing treatments on a particular criterion. Then, lower and upper bounds of evidential support (termed belief and plausibility) are calculated per hypothesis, for example ‘the evidential support for this hypothesis is between 50% and 80%’. Finally, probability mass assignments from different evidence sources (for example, different decision makers) can be combined with Dempster’s rule of combination. An agreement metric between decision makers can also be calculated.

Fuzzy set theory

In fuzzy set theory, elements have a degree of membership to a set [35,36]. The degree of an element’s membership to a fuzzy set is expressed as a number between zero (not a member of the fuzzy set) and one (completely a member of the fuzzy set). Degrees of memberships between zero and one indicate ambiguous set membership. Consider as examples of fuzzy sets the sets of “very important criteria” or of “low criteria weights”. If all memberships are equal to either zero or one, fuzzy set theory reduces to conventional set theory.

When applying fuzzy set theory for MCDA, decision makers first have to identify ambiguous elements (such as particular criteria weights) in their decision problem. They then have to define fuzzy sets and the membership functions to capture the identified ambiguity. For example, the pairwise comparisons to establish criteria weights in the AHP are conventionally numbers between 1 (“equivalence”) and 9 (“extreme preference”) [37]. When decision makers use AHP in combination with fuzzy sets, statements such as “extreme preference” can be ambiguous. The ambiguity in such statements can be represented with fuzzy sets [38,39].

Grey theory

In Grey theory, uncertainty can be represented with ranges termed black, white or grey numbers [40]. The ‘shade’ of a number indicates the magnitude of uncertainty. Black numbers represent a complete lack of knowledge (range is from minus infinity to plus infinity), whereas white numbers represent complete knowledge (range is a single number). Grey numbers are between these extremes, for example a grey number with a lower bound of 1 and an upper bounds of 5. Like fuzzy sets, grey numbers can be described by verbal statements: for example, a performance score between 0 and 0.3 on a particular criterion may be defined as “low” [41]. In an MCDA context, grey theory requires decision makers to provide lower and upper bounds for criteria weights or performance scores. These yield bounds on the overall treatment scores.

MCDCA Method	Bayesian framework	DSA	Fuzzy set theory	Grey theory	PSA	n(%)
Outranking						
*DRSA[77]	-	-	-	-	2	2 (0%)
*ELECTRE[72]	1	10	9	-	3	23 (3%)
*PROMETHEE[72]	1	17	14	1	17	50 (7%)
Value-based						
*AHP[72]	18	116	174	6	34	348 (52%)
ANP[72]	-	7	10	-	-	17 (3%)
*MACBETH[72]	1	1	-	-	1	3 (0%)
*MAUT[72]	1	8	-	-	5	14 (2%)
*MAVT[2]	-	6	-	-	5	11 (2%)
OWA[78]	3	1	12	-	2	18 (3%)
SAW[78]	-	4	2	1	2	9 (1%)
SMAA[66]	-	-	-	-	10	10 (1%)
*SMART[79]	-	2	-	-	1	3 (0%)
WSM[78]	-	4	-	-	3	7 (1%)
Reference-based						
ER[77]	7	4	17	5	6	39 (6%)
TOPSIS[72]	-	6	45	2	4	57 (9%)
VIKOR[80]	-	2	5	-	-	7 (1%)
Other/hybrid						
DMCE[81]	-	2	1	-	-	3 (0%)
TODIM[82]	-	1	1	-	-	2 (0%)
Other	5	9	16	8	8	46 (7%)
<hr/>						
n(%)	37 (6%)	200 (30%)	306 (46%)	23 (3%)	103 (15%)	669 (100%)

Table 3 (on previous page): This table shows how often the identified uncertainty approaches were combined with existing MCDA methods, the total number of abstracts per approach and the total number of abstracts per MCDA method. A single reference to a relevant textbook is added next to each one of the MCDA method names (if available, else a paper is cited). Note that it was possible for studies to apply more than one MCDA method. Some studies applied more than one MCDA method and were counted for each method. *the name of this MCDA method was used in the search strategy. The (English) meanings of the MCDA method abbreviations are as follows. DRSA=Dominance-based rough set approach. ELECTRE=Elimination and choice translating reality method. PROMETHEE=Preference ranking organization method for enrichment evaluation. AHP=Analytic hierarchy process. ANP=Analytic network process. MACBETH=Measuring attractiveness by a categorical-based evaluation technique. MAUT=Multi-attribute utility theory. MAVT=Multi-attribute value theory. OWA=Ordered weighted average. SAW=Simple additive weighting. SMAA=Stochastic multi-criteria acceptability analysis. SMART=Simple multi-attribute rating technique. WSM=Weighted sum method. ER=evidential reasoning. TOPSIS=Technique for order preference by similarity to an ideal solution. VIKOR=Multicriteria Optimization and Compromise Solution. DMCE=deliberative multi-criteria evaluation. TODIM=Interactive and multicriteria decision making. DSA=deterministic sensitivity analysis. PSA=probabilistic sensitivity analysis

Applications in healthcare

Nineteen applications of the approaches in healthcare-related publications were identified. Of these, seven are related to healthcare policy decisions. Nine studies in healthcare used the deterministic approach [5,42–49]. Of these, four studies were in the context of (research) portfolio optimization [42,46,47,49], early health technology assessment [46], and benefit-risk assessment [5]. The other studies applying the deterministic approach were in emergency management [43] and drinking water systems [44,45]. Four studies in healthcare were categorized as probabilistic [9,23,50,51], of which two focus on benefit-risk assessments [9,51], one on infectious diseases [23] and one on water transport (safety) [50]. Four fuzzy set theory studies considered environmental health issues [52–55], while one considered diagnostics [56]. Finally, one study applied a Bayesian framework for diagnostics [57].

Discussion

Comparison of approaches

The present review was performed to identify and classify the different approaches to quantify uncertainty in MCDA. Five distinct approaches were identified: deterministic sensitivity analysis, probabilistic sensitivity analysis, Bayesian framework, fuzzy set theory and grey theory. To guide our discussion on the advantages and disadvantages of these approaches for healthcare applications, we will discuss them with respect to six criteria that are derived from earlier studies that assessed the applicability of uncertainty approaches for: operations research [58], forestry [18], engineering [59], and health economic models [60]. Criteria relating to what extent approaches can represent uncertainty are inputs and

outputs (how can decision makers assign uncertainty to model parameters and what additional information does the approach then yield) and the number of uncertainty sources that can be taken into account. More practical considerations are the versatility of the approaches with regard to combining them with MCDA methods, time considerations, and prerequisite knowledge.

Deterministic sensitivity analysis implies that weights are varied as a single value and is therefore easy applicable to both uncertainty in performance scores [20] and uncertainty in criteria weights [61,62]. If there is heterogeneity in criteria performance scores and/or criteria weights, scenario analysis can be used to compare the outcomes for relevant subgroups [46,63]. Drawbacks of deterministic sensitivity analysis are that the range over which weights or performance scores are varied is usually arbitrarily chosen and that it is assumed that all parameter values in the range are equally probable. These drawbacks may lead to a biased view of the impact of uncertainty on the decision. Probabilistic sensitivity analysis can address these particular drawbacks by allowing decision makers to assign probability distributions estimating both stochastic and parameter uncertainty.

Bayesian networks are especially relevant when there are conditional relations in the evidence sources which obviously is present if confounded clinical endpoints obtained from clinical studies are transformed to a performance score. Bayesian networks seem therefore mostly useful as a method to investigate the evidence before the scoring step of alternatives in an MCDA. Dempster-Shafer theory is most useful when little or no evidence is available and an elicitation method is used to gather expert opinion on the performance scores of treatments. Because human judgment is often characterized by *ambiguity* [35,64], decision makers may accept fuzzy set theory more readily than approaches that denote variation in criteria weights as *uncertain* using terms like ‘deviation’ and ‘error’ [35]. In grey theory, ranges can be defined for both criteria weights and performance scores easily. However, this gives information only about the bounds of model parameters such as overall treatment scores and does not give insight into the likelihood of values in between the bounds.

Widening the application of uncertainty analysis in MCDA for healthcare

In an attempt to develop guidance for practitioners of MCDA, the five approaches can be compared in terms of how the required input is elicited and what additional information the approaches yield about the magnitude and impact of uncertainty. Three input modes

are defined: changing values, specifying ranges and specifying distributions. In “changing values”, decision makers simply take other values for the criteria weights and performance scores. Although theoretically this can be done with any approach, it is what deterministic sensitivity analysis was designed specifically for. The output of such an analysis can give decision makers more insight in the impact of the induced variation on the overall treatment scores and on the rank order of treatments. In “specifying ranges”, decision makers have to specify lower and upper bounds of model parameters. This can be done with the probabilistic, grey and fuzzy set approaches. Grey theory was specifically designed for this input mode. In probabilistic sensitivity analysis, a uniform distribution between the lower and upper bounds can be assigned. In fuzzy set theory, uniform fuzzy sets can be defined between the lower and upper bounds. The outputs yielded by the approaches differ when using the “specifying ranges” input mode. Grey theory approach will only provide insight into bounds for treatment overall scores while the probabilistic and fuzzy approaches also yield the distribution of treatment overall scores between the bounds. Finally, by “specifying distributions”, decision makers state how values are distributed over a range. Distributions can be specified in probabilistic sensitivity analysis and fuzzy set theory. As output the decision makers will gain insight into the distribution of overall alternative scores. The overlap between the distributions of treatment overall scores can be used to assess the impact of uncertainty on the rank order of treatments. In the probabilistic and Bayesian approaches, this is operationalized as the probability of particular rankings occurring.

Deterministic sensitivity analysis is the only approach that cannot take into account a larger number of uncertain model parameters simultaneously and thus does not consider the cumulative impact of uncertainty in multiple model parameters. The probabilistic and Bayesian network approaches can simultaneously take into account uncertainty from multiple sources of uncertainty with Monte Carlo simulations. In Dempster-Shafer theory, Dempster’s rule of combination is used to combine the probability mass assessments of multiple decision makers. Fuzzy sets and grey numbers can be combined using the known mathematical operations on sets and ranges [35,65].

In this review, approaches to take into account uncertainty are identified and classified according to their ability to capture and represent uncertainty in the elicitation of criteria weights and performance scores. However, the applicability of these approaches are sometimes strictly dependent on the specific form of MCDA used. Some MCDA methods such as stochastic multi-criteria acceptability analysis are very closely tied to one specific uncertainty approach (in this case, the probabilistic approach). Some other MCDA methods

allow the use of multiple approaches for uncertainty while AHP and PROMETHEE can be used with all approaches. It is yet unclear whether other gaps in combinations are due to fundamental methodological mismatches or if the combinations are theoretically possible when there is more familiarity with the MCDA methods and/or uncertainty approaches.

With regard to time considerations, little time is required for simple deterministic sensitivity analysis (assuming only one or two parameters are changed simultaneously). The process of assigning probability distributions in the probabilistic approach can be time-consuming for analysts (when a large amount of data has to be modelled) and decision makers (when distributions are elicited from them). This duration can be reduced by assigning distributions only to specific parameters. An example of when this is relevant is when clinical data is available but decision makers are unable or unwilling to provide criteria weights [9,27,51,66]. Time requirements in Bayesian framework are more demanding because of the assignment of not only probability distributions but also of dependence relations in the form of conditional probabilities. Fuzzy set theory requires the definition of fuzzy sets which takes time. Yet, when these are agreed upon they can be used over multiple decisions. Grey theory is straightforward to use in a group discussion setting [67]. If there is disagreement about the value of model parameters in the group, the lowest and highest of those can be used as bounds for the Grey number.

A final practical consideration is the knowledge required to implement an uncertainty approach or to interpret its results. Deterministic sensitivity analysis requires no additional knowledge apart from knowledge about the MCDA method that is used. Grey theory requires decision makers to be able give define and interpret ranges of values. Probabilistic sensitivity analysis requires that decision makers are familiar with probability distributions. This is also the case with Bayesian frameworks, which in addition require knowledge about Bayesian statistics. Analysts applying the Bayesian framework need knowledge about Bayesian programming languages such as WINBUGS. Decision makers should be familiar with set theory to be able to understand and apply fuzzy set theory. When there is a disparity between the decision maker's current knowledge and the knowledge required from the approach, there is a knowledge gap. This gap may lead to a lack of confidence in the results of an (uncertainty) analysis [68,69], and bridging the gap can be time-consuming. Apart from the required knowledge, visual representations of uncertainty are important factors for ease of interpretation. Decision makers applying deterministic sensitivity analysis can obtain a tornado graph which ranks model parameters on their ability to change the overall scores of alternatives. For probabilistic and Bayesian approaches, scatter plots or density plots can be used and Bayesian networks can also be

shown. Fuzzy sets can be visualized through membership function plots similar to probabilistic density plots. Since the outcomes of Dempster-Shafer theory and grey theory analysis are lower and upper bounds of overall scores of treatments, graphs may be less useful.

Limitations

Although the present review identified many applications, the list is unlikely to be exhaustive due to the large amount of work on MCDA in different fields. Furthermore, studies that did not mention uncertainty in their title or abstract may have been missed. Although these are potential limitations, the sample of studies provides sufficient information to stimulate a discussion about the use of approaches for uncertainty assessment. Approaches that our review did not classify as such but that were mentioned in the earlier reviews are risk-based approaches [17] and possibility theory [18]. In risk-based approaches, an alternative's performance score on a criterion will become lower when that performance is uncertain ('more risky'). Possibility theory combines fuzzy set theory and evidential reasoning [70]. Although these approaches were classified as distinct approaches in the earlier reviews, there is considerable overlap with our classification.

The aspects on which we compared the uncertainty approaches are based on earlier literature, yet it is important to acknowledge that for real-world decision making other aspects, depending on the specific decision and decision maker(s), may be relevant. Further empirical research with decision makers is needed to better assess the usefulness and specific requirements of the approaches for real-world decision making.

Following the classification of sources of uncertainty, all approaches that were identified can be used to assess uncertainty in the criteria weights and performance scores as assigned by decision makers. However, no approaches were identified to deal with structural uncertainty [2,71–73]. One explanation for this is that MCDA already facilitates an informed discussion and that this addresses structural uncertainty. However, further work is recommended to identify approaches to take into account structural uncertainty and ways to develop MCDA models to incorporate these approaches [74,75].

Conclusions and recommendations

To our best knowledge, our review is the first to give an overview of approaches to take into account uncertainty in MCDA-supported decisions with a focus on the approaches'

applicability to the context of healthcare decision making. The review identified five approaches to take into account uncertainty in MCDA-supported decisions. In conclusion, the deterministic approach seems most appropriate if the criteria weights or performance scores are varied as a single value. The grey approach seems most appropriate if only lower and upper bounds are elicited. The other approaches can be used flexibly across all three input modes. In a group decision process where the opinions of several decision makers are combined, the distribution input mode seems most relevant, and this would argue for the probabilistic and fuzzy set approaches that allow distributions [76]. From the review it is concluded that the deterministic sensitivity analysis will likely be sufficient for most decisions and has the advantage of a straightforward implementation. It also is an intuitive approach and applicable for multiple MCDA approaches. Although deterministic sensitivity analysis is useful in many applications, the more complex approaches are especially useful when uncertainty in multiple parameters has to be taken into account simultaneously, when dependence relations exist, or when there are no prohibitive time constraints for uncertainty modeling.

References

1. Diaby V, Campbell K, Goeree R. Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis. *Oper. Res. Heal. Care.* 2013;2:20–4.
2. Belton V, Stewart TJ. Multiple criteria decision analysis : An integrated approach. 2nd ed. Dordrecht: Kluwer Academic Publishers; 2002.
3. Hummel JM, Bridges JFP, IJzerman MJ. Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. *Patient.* 2014;7:129–40.
4. Holden W. Benefit-risk analysis: A brief review and proposed quantitative approaches. *Drug Saf.* 2003;26:853–62.
5. Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines — part 1 : the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiol. Drug Saf.* 2007;16:S2–15.
6. Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. *Value Heal.* 2010;13:657–66.
7. Phillips LD, Fasolo B, Zafiroopoulos N, Beyer A. Is quantitative benefit–risk modelling of drugs desirable or possible? *Drug Discov. Today Technol.* 2011;8:e3–10.
8. Levitan BS, Andrews EB, Gilsenan A, Ferguson J, Noel R a, Coplan PM, et al. Application of the BRAT framework to case studies: observations and insights. *Clin. Pharmacol. Ther.* 2011;89:217–24.
9. Van Valkenhoef G, Tervonen T, Zhao J, de Brock B, Hillege HL, Postmus D. Multicriteria benefit-risk assessment using network meta-analysis. *J. Clin. Epidemiol.* 2012;65:394–403.

10. Diaby V, Goeree R. How to use multi-criteria decision analysis methods for reimbursement decision-making in healthcare: a step-by-step guide. *Expert Rev. Pharmacoecon. Outcomes Res.* 2014;14:81–99.
11. Dolan JG. Shared decision-making--transferring research into practice: the Analytic Hierarchy Process (AHP). *Patient Educ. Couns.* 2008;73:418–25.
12. Thokala P, Duenas A. Multiple Criteria Decision Analysis for Health Technology Assessment. *Value Heal. Elsevier*; 2012;15:1172–81.
13. Devlin NJ, Sussex JON. Incorporating multiple criteria in HTA: methods and processes. London; 2011.
14. Bojke L, Claxton K, Sculpher M, Palmer S. Characterizing structural uncertainty in decision analytic models: a review and application of methods. *Value Heal.* 2009;12:739–49.
15. Briggs AH, Weinstein MC, Fenwick E a L, Karnon J, Sculpher MJ, Paltiel a D. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Heal.* 2012;15:835–42.
16. Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the Value of Healthcare Interventions Using Multi-Criteria Decision Analysis: A Review of the Literature. *Pharmacoeconomics.* 2014;1–21.
17. Durbach IN, Stewart TJ. Modeling uncertainty in multi-criteria decision analysis. *Eur. J. Oper. Res. Elsevier B.V.*; 2012;223:1–14.
18. Kangas AS, Kangas J. Probability, possibility and evidence: approaches to consider risk and uncertainty in forestry decision analysis. *For. Policy Econ.* 2004;6:169–88.
19. Felli J, Hazen GB. Sensitivity analysis and the expected value of perfect information. *Med. Decis. Mak.* 1998;18:95–109.
20. Churilov L, Liu D, Ma H, Christensen S, Nagakane Y, Campbell B, et al. Multiattribute selection of acute stroke imaging software platform for Extending the Time for Thrombolysis in Emergency Neurological Deficits (EXTEND) clinical trial. *Int. J. Stroke.* 2013;8:204–10.
21. Parthiban P, Abdul Zubar H. An integrated multi-objective decision making process for the performance evaluation of the vendors. *Int. J. Prod. Res.* 2013;51:3836–48.
22. Hyde KM, Maier HR, Colby CB. A distance-based uncertainty analysis approach to multi-criteria decision analysis for water resource decision making. *J. Environ. Manage.* 2005;77:278–90.
23. Cox R, Sanchez J, Revie CW. Multi-criteria decision analysis tools for prioritising emerging or re-emerging infectious diseases associated with climate change in Canada. *PLoS One.* 2013;8:e68338.
24. Hopfe CJ, Augenbroe GLM, Hensen JLM. Multi-criteria decision making under uncertainty in building performance assessment. *Build. Environ.* 2013;69:81–90.
25. Haakma W, Steuten LMG, Bojke L, IJzerman MJ. Belief elicitation to populate health economic models of medical diagnostic devices in development. *Appl. Health Econ. Health Policy.* 2014;12:327–34.
26. Betrie GD, Sadiq R, Morin KA, Tesfamariam S. Selection of remedial alternatives for mine sites: a multicriteria decision analysis approach. *J. Environ. Manage.* 2013;119:36–46.
27. Prado-Lopez V, Seager TP, Chester M, Laurin L, Bernardo M, Tylock S. Stochastic multi-attribute analysis (SMAA) as an interpretation method for comparative life-cycle assessment (LCA). *Int. J. Life Cycle Assess.* 2013;19:405–16.
28. Broekhuizen H, Groothuis-Oudshoorn C, Hauber A, IJzerman M. Integrating patient preferences and clinical trial data in a Bayesian model for benefit-risk assessment. 34th Annu. Meet. Soc. Med. Decis. Mak. Phoenix, AZ; 2012.
29. Fenton N, Neil M. Making decisions: using Bayesian nets and MCDA. *Knowledge-Based Syst.* 2001;14:307–25.

30. Albert J. Bayesian computation with R. 1st ed. Gentleman R, Hornik K, Parmigiani G, editors. New York city: Springer Science + Business Media, Inc.; 2007.
31. Beynon MJ. The Role of the DS/AHP in Identifying Inter-Group Alliances and Majority Rule Within Group Decision Making. *Gr. Decis. Negot.* 2005;15:21–42.
32. Srivastava RP. An introduction to evidential reasoning for decision making under uncertainty: Bayesian and belief function perspectives. *Int. J. Account. Inf. Syst.* Elsevier Inc.; 2011;12:126–35.
33. Ma W, Xiong W, Luo X. A Model for Decision Making with Missing , Imprecise , and Uncertain Evaluations of Multiple Criteria. *Int. J. Intell. Syst.* 2013;28:152–84.
34. Beynon M, Curry B, Morgan P. The Dempster-Shafer theory of evidence: an alternative approach to multicriteria decision modelling. *Omega.* 2000;28:37–50.
35. Zadeh L. Fuzzy sets. *Inf. Control.* 1965;353:338–53.
36. Chen S-J, Hwang C-L. Fuzzy multiple attribute decision making. 1st ed. Berlin: Springer-Verlag; 1992.
37. Saaty RW. The Analytical Hierarchy Process - What It Is And How It Is Used. *Math. Model.* 1987;9:161–76.
38. Montazar A, Gheidari ON, Snyder RL. A fuzzy analytical hierarchy methodology for the performance assessment of irrigation projects. *Agric. Water Manag.* 2013;121:113–23.
39. Pitchipoo P, Venkumar P, Rajakarunakaran S. Fuzzy hybrid decision model for supplier evaluation and selection. *Int. J. Prod. Res.* 2013;51:3903–19.
40. Ju-Long D. Control problems of grey systems. *Syst. Control Lett.* 1982;1:288–94.
41. Kuang H, Hipel KW, Kilgour DM. Evaluation of source water protection strategies in Waterloo Region based on Grey Systems Theory and PROMETHEE II. 2012 IEEE Int. Conf. Syst. Man, Cybern. IEEE; 2012. p. 2775–9.
42. Franc M, Frangiosa T. Pharmacogenomic study feasibility assessment and pharmaceutical business decision-making. *Methods Pharmacol. Toxicol.* 2008. p. 253–65.
43. Bertsch V, Geldermann J. Preference elicitation and sensitivity analysis in multicriteria group decision support for industrial risk and emergency management. *Int. J. Emerg. Manag.* 2008;5:7.
44. Bouchard C, Abi-Zeid I, Beauchamp N, Lamontagne L, Desrosiers J, Rodriguez M. Multicriteria decision analysis for the selection of a small drinking water treatment system. *J. Water Supply Res. Technol.* 2010;59:230.
45. Francisque A, Rodriguez MJ, Sadiq R, Miranda LF, Proulx F. Reconciling “actual” risk with “perceived” risk for distributed water quality: a QFD-based approach. *J. Water Supply Res. Technol.* 2011;60:321.
46. Hilgerink MP, Hummel MJ, Manohar S, Vaartjes SR, Ijzerman MJ. Assessment of the added value of the Twente Photoacoustic Mammoscope in breast cancer diagnosis. *Med. Devices (Auckl).* 2011;4:107–15.
47. Lee CW, Kwak NK. Strategic enterprise resource planning in a health-care system using a multicriteria decision-making model. *J. Med. Syst.* 2011;35:265–75.
48. Richman MB, Forman EH, Bayazit Y, Einstein DB, Resnick MI, Stovsky MD. A novel computer based expert decision making model for prostate cancer disease management. *J. Urol.* 2005;174:2310–8.
49. Wilson E, Sussex J, Macleod C, Fordham R. Prioritizing health technologies in a Primary Care Trust. *J. Health Serv. Res. Policy.* 2007;12:80–5.
50. Rafiee R, Ataei M, Kamali M. Probabilistic stability analysis of Naien water transporting tunnel and selection of support system using TOPSIS approach. *Sci. Res. Essays.* 2011;6:4442–54.

51. Tervonen T, van Valkenhoef G, Buskens E, Hillege HL, Postmus D. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Stat. Med.* 2011;30:1419–28.
52. Karimi AR, Mehrdadi N, Hashemian SJ, Nabi Bidhendi GR, Tavakkoli Moghaddam R. Selection of wastewater treatment process based on the analytical hierarchy process and fuzzy analytical hierarchy process methods. *Int. J. Environ. Sci. Technol.* 2011;8:267–80.
53. Tesfamariam S, Sadiq R, Najjaran H. Decision making under uncertainty--an example for seismic risk management. *Risk Anal.* 2010;30:78–94.
54. Yang AL, Huang GH, Qin XS, Fan YR. Evaluation of remedial options for a benzene-contaminated site through a simulation-based fuzzy-MCDA approach. *J. Hazard. Mater.* 2012;213-214:421–33.
55. Zhang K, Achari G, Pei Y. Incorporating linguistic, probabilistic, and possibilistic information in a risk-based approach for ranking contaminated sites. *Integr. Environ. Assess. Manag.* 2010;6:711–24.
56. Uzoka FME, Barker K. Expert systems and uncertainty in medical diagnosis: a proposal for fuzzy-ANP hybridisation. *Int. J. Med. Eng. Inform.* 2010;2:329.
57. Castro F, Caccamo LP, Carter KJ, Erickson BA, Johnson W, Kessler E, et al. Sequential Test Selection in the Analysis of Abdominal Pain. *Med. Decis. Mak.* 1996;16:178–83.
58. Rios Insua D, French S. A framework for sensitivity analysis in discrete multi-objective decision-making. *Eur. J. Oper. Res.* 1991;54:176–90.
59. Helton J. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliab. Eng. Syst. Saf.* 1993;42:327–67.
60. Claxton K. Briefing paper for methods review workshop on exploring uncertainty. 2007 p. 1–63.
61. Ribeiro F, Ferreira P, Araújo M. Evaluating future scenarios for the power generation sector using a Multi-Criteria Decision Analysis (MCDA) tool: The Portuguese case. *Energy.* 2013;52:126–36.
62. Srdjevic Z, Samardzic M, Srdjevic B. Robustness of AHP in selecting wastewater treatment method for the coloured metal industry: Serbian case study. *Civ. Eng. Environ. Syst.* 2012;29:147–61.
63. Durbach IN. Outranking under uncertainty using scenarios. *Eur. J. Oper. Res.* 2014;232:98–108.
64. Vahdani B, Mousavi SM, Tavakkoli-Moghaddam R, Hashemi H. A new design of the elimination and choice translating reality method for multi-criteria group decision-making in an intuitionistic fuzzy environment. *Appl. Math. Model.* 2013;37:1781–99.
65. Yang Y. Extended grey numbers and their operations. 2007 IEEE Int. Conf. Syst. Man Cybern. Ieee; 2007;2181–6.
66. Lahdelma R, Hokkanen J. SMAA-stochastic multiobjective acceptability analysis. *Eur. J. Oper.* 1998;2217:137–43.
67. Zavadskas EK, Kaklauskas A. Multi-Attribute Decision-Making Model by Applying Grey Numbers. *Informatica.* 2009;20:305–20.
68. Caro J, Möller J. Decision-analytic models: current methodological challenges. *Pharmacoeconomics.* 2014;
69. Tuffaha HW, Gordon LG, Scuffham P a. Value of Information Analysis in Health Care: A Review of Principles and Applications. *J. Med. Econ.* 2014;1–7.
70. Zadeh LA. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* 1978;1:3–28.
71. Keeney R, Raiffa H. Decisions with multiple objectives. Cambridge: Cambridge University Press; 1976.
72. Ishizaka A, Nemery P. Multi-criteria decision analysis: methods and software. 1st ed. John Wiley & Sons Ltd; 2013.

73. Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. *Med. Decis. Mak.* 2011;31:675–92.
74. Winterfeldt D. Structuring decision problems for decision analysis. *Acta Psychol. (Amst).* 1980;45:71–93.
75. Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. *Econometrica.* 1979;47:263–91.
76. Belton V, Pictet J. A Framework for Group Decision Using a MCDA Model: Sharing, Aggregating or Comparing Individual Information? *J. Decis. Syst.* 1997;6:283–303.
77. Liu W. Propositional, probabilistic and evidential reasoning. Heidelberg: Physica-Verlag HD; 2001.
78. Bouyssou D, Marchant T, Pirlot M, Tsoukias A, Vincke P. Evaluation and decision models with multiple criteria: stepping stones for the analyst. New York: Springer Science + Business Media, Inc.; 2006.
79. Winterfeldt D, Edwards W. Decision analysis and behavioral research. Cambridge: Cambridge University Press; 1986.
80. Opricovic S. Multicriteria optimization of civil engineering systems. Belgrade: Faculty of Civil Engineering; 1998.
81. Proctor W, Drechsler M. Deliberative multicriteria evaluation. *Environ. Plan. C Gov. Policy.* 2006;24:169–90.
82. Autran Monteiro Gomes LF, Duncan Rangel LA. An application of the TODIM method to the multicriteria rental evaluation of residential properties. *Eur. J. Oper. Res.* 2009;193:204–11.

Chapter 3

Estimating the value of medical treatments to patients
using probabilistic multi criteria decision analysis

Henk Broekhuizen

Catharina G.M. Groothuis-Oudshoorn

A. Brett Hauber

Jeroen P. Jansen

Maarten J. IJzerman

Published as: Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber AB, Jansen JP, IJzerman MJ. Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis. BMC Medical Informatics and Decision Making 2015;15:1–10.

Abstract

Background: Estimating the value of medical treatments to patients is an essential part of healthcare decision making, but is mostly done implicitly and without consulting patients. Multi criteria decision analysis (MCDA) has been proposed for the valuation task, while stated preference studies are increasingly used to measure patient preferences. In this study we propose a methodology for using stated preferences to weigh clinical evidence in an MCDA model that includes uncertainty in both patient preferences and clinical evidence explicitly.

Methods: A probabilistic MCDA model with an additive value function was developed and illustrated using a case on hypothetical treatments for depression. The patient-weighted values were approximated with Monte Carlo simulations and compared to expert-weighted results. Decision uncertainty was calculated as the probability of rank reversal for the first rank. Furthermore, scenario analyses were done to assess the relative impact of uncertainty in preferences and clinical evidence, and of assuming uniform preference distributions.

Results: The patient-weighted values for drug A, drug B, drug C, and placebo were 0.51 (95% CI: 0.48 to 0.54), 0.51 (95% CI: 0.48 to 0.54), 0.54 (0.49 to 0.58), and 0.15 (95% CI: 0.13 to 0.17), respectively. Drug C was the most preferred treatment and the rank reversal probability for first rank was 27%. This probability decreased to 18% when uncertainty in performances was not included and increased to 41% when uncertainty in criterion weights was not included. With uniform preference distributions, the first rank reversal probability increased to 61%. The expert-weighted values for drug A, drug B, drug C, and placebo were 0.67 (95% CI: 0.65 to 0.68), 0.57 (95% CI: 0.56 to 0.59), 0.67 (95% CI: 0.61 to 0.71), and 0.19 (95% CI: 0.17 to 0.21). The rank reversal probability for the first rank according to experts was 49%.

Conclusions: Preferences elicited from patients can be used to weigh clinical evidence in a probabilistic MCDA model. The resulting treatment values can be contrasted to results from experts, and the impact of uncertainty can be quantified using rank probabilities. Future research should focus on integrating the model with regulatory decision frameworks and on including other types of uncertainty.

Introduction

Decisions in healthcare policy regarding research portfolio management, market access, reimbursement and price-setting all depend (in part) on the added value of medical treatments for patients. This treatment valuation task is difficult because it has to be based on a large set of (possibly uncertain) clinical evidence and on subjective assessments of the desirability of clinical endpoints. Multi criteria decision analysis (MCDA) is a decision analytic modelling approach that may be useful for such treatment valuation tasks [1,2], primarily because it can support decision makers by structuring the available evidence [3,4] and by guiding informed discussions through visualizations [5]. In MCDA, the decision goal (in our case, valuing treatments) is decomposed into a set of concrete and measurable *criteria* (in our case, clinical endpoints or treatment characteristics like mode of administration). The identification of this set of criteria can be done, for example, by interviewing patients and clinical experts. Then, the set of relevant decision options (termed *alternatives*) is defined. These are often a given in a treatment valuation task. Now that the structure of the MCDA model is built, two main inputs are required: criterion weights and performance scores. Criterion weights indicate the relative importance of criteria. Performance scores measure the experts' assessment of how well the alternatives perform on each of the criteria. Criterion weights and performance scores can be aggregated to come to an overall value of each included treatment [6]. This overall value can then be used to select a most preferred treatment, to rank treatments from best to worst, or to sort treatments into categories.

Studies applying MCDA to the treatment valuation task can, for example, be found in the decision contexts of market access [7–9] and reimbursement [10–12]. These applications of MCDA have mostly used expert input to construct the criterion weights and performances scores. However, it has been argued that the patient perspective forms an essential part of treatment value [13–16]. In an MCDA framework this could be operationalized by letting patients set the criterion weights. One approach for this is to involve individual patient representatives in the decision making process, but a more representative approach would be to use stated preference methods to elicit preferences from a large group of patients [17,18]. These patient preferences could then be used to weigh the available clinical evidence [19]. In that way, treatment value can be estimated from the patient's perspective in a transparent and representative manner. The results from such analyses could then be used as input for the decision makers' decision making process.

In its simplest form, this combination of patient preferences with clinical evidence can be done deterministically. This would imply that the mean criterion weights and mean performance scores are used as input for the MCDA. However, including an assessment of uncertainty in a decision analysis would be advantageous because 1) it can help assess confidence in the outcomes of the model, 2) it can help ascertain the usefulness of performing additional research [20], and 3) can prevent bias in non-linear models [21]. Several approaches exist for taking into account uncertainty in MCDA. A recent review into uncertainty analysis approaches that are potentially useful for the specific context of healthcare identified deterministic sensitivity analysis, probabilistic sensitivity analysis, Bayesian frameworks, grey theory and fuzzy set theory [22]. The review concluded that deterministic sensitivity analysis is likely sufficient for most decisions in healthcare but that for decisions where the views of multiple stakeholders are combined or when uncertainty in multiple parameters is to be considered simultaneously, approaches that allow distributions (such as the probabilistic approach) would be more appropriate. The treatment valuation task considered in this paper requires the combination of the views from multiple stakeholders (namely, a large group of patients) and requires the combination of uncertainty in multiple parameters (namely, all weights and performance scores), and the probabilistic approach is therefore adopted in this study.

The aim of this study is to illustrate how patients' criterion weights derived from a stated preference study together with performance scores derived from clinical evidence can be used to value treatments from the patient's perspective, taking into account parameter uncertainty in both criterion weights and performance scores. A hypothetical case based on earlier studies concerning three antidepressants and placebo will be presented to illustrate the developed model. Its main outputs are patient-weighted treatment values with associated 95% confidence intervals. It will be shown how the patient valuation can be contrasted to an expert-based valuation and the utility of the developed modeling approach for practical decision making will be further illustrated by present the results from three scenario analyses.

Model parameters and notations

Suppose I treatments have to be valued based on n criteria simultaneously. We define treatments with a higher value to be preferred to treatments with a lower value. The clinical performance of drug i on criterion k is denoted with θ_{ik} . The partial value function $v_k(\theta)$ for criterion k maps the criterion-specific performance values θ_{ik} onto a linear scale

between 0 at a ‘worst imaginable’ performance of θ_k^- and 1 at a ‘best imaginable’ performance of θ_k^+ for treatment i :

$$v_k(\theta_{ik}) = \begin{cases} 1, & \text{if } \theta_{ik} \geq \theta_k^+ \\ \frac{\theta_{ik} - \theta_k^-}{\theta_k^+ - \theta_k^-}, & \text{if } \theta_k^- < \theta_{ik} < \theta_k^+ \\ 0, & \text{if } \theta_{ik} \leq \theta_k^- \end{cases} \quad (1)$$

The weights of the criteria are denoted with w_k . These criterion weights indicate the relative importance of scale swings from θ_k^- to θ_k^+ on a criterion, and should be estimated using the swing direct method or the MACBETH pairwise comparisons method [4]. To come to an overall value V_i for each treatment i , the partial values are combined with an additive value function

$$V_i = \sum_{k=1}^n w_k v_k(\theta_{ik}), \quad (2)$$

where it is assumed that the criteria are independent.

Taking into account uncertainty

We adopt a probabilistic framework, in which the uncertainty in estimates for criterion weight and performance scores are represented with probability distributions [22]. The partial value functions and the overall values are stochastic variables with probability distributions that are complex combinations of the probability distributions for the weights and treatment performances. These are hard to calculate analytically and will therefore be approximated by applying a Monte Carlo simulation approach. In such an approach, for each simulation run t , weights w_{kt} and performances θ_{ikt} are sampled from their respective probability distributions. Then, formula's 1 and 2 are used to come to partial values $v_k(\theta_{ikt})$ and overall values V_{it} . This process is then repeated a large number of times T .

The main outcomes of the MCDA model are the mean overall value for each treatment, the value distributions for each treatment, and the ranking probabilities for each treatment. The mean overall value for treatment i is estimated with the posterior mean, that is $V_i = \frac{\sum V_{it}}{T}$. The value distribution of treatment i is the empirical distribution of all v_{it} . Rank probabilities are calculated by ranking treatments in descending order on their overall value each Monte Carlo simulation run. We define r_{xi} as the amount of Monte Carlo simulation runs where treatment i attains rank x . Then, treatment i 's rank probability for

rank x is $\frac{r_{xi}}{T}$. The probability that the treatment with the highest mean value is not ranked first is used as a measure of decision uncertainty. It is calculated as follows: terming treatment j the treatment with the highest mean value, the probability that this treatment is not ranked first is $1 - \frac{r_{1j}}{T}$.

Illustration using a case

The model is illustrated with a case on treatments for severe depression. As can be seen in Figure 1, the included treatments are compared on four criteria: response, remission, adverse events and severe adverse events. Response is defined as the probability of a 50% reduction in acute depression symptoms as measured on a depression scale such as the Hamilton rating scale for depression or the Montgomery Asberg depression rating scale. Remission is defined as the probability that depressive symptoms are reduced for such a time such that a patient can be considered to have recovered from an acute depressive episode. Adverse events considered are (probability of) sexual dysfunction, hypertension, restlessness, sedation, dizziness, nausea, dry mouth, sweating and weight increase. Severe adverse events considered are (probability of) suicide and other events that lead to death, threat to life, permanent/severe disability or hospitalization.

The response and remission criteria are to be maximized, with the best level θ^+ defined as 100% and the worst level θ^- as 0%. The adverse events and severe adverse events are to be minimized, with the best level θ^+ defined as 0% and the worst level θ^- as 100%. This means that, if for a patient the weight for response is 0.4 and the weight for remission is 0.2, that patient considers the swing from 0% probability of response to 100% probability of response to be twice as important as the swing from 0% probability of remission to 100% probability of remission when choosing an antidepressant.

For illustration purposes hypothetical preference and performance datasets were used. A patient criterion weight sample of 100 patients was constructed by bootstrapping the results from a patient panel held in an earlier elicitation study [23]. In that study, weights were also elicited from five clinical experts. These are included in our study for comparison with patient preferences. Three hypothetical antidepressants and placebo are included. We define drugs “A” and “B” as the currently used drugs. They are assumed to have moderate effectiveness and side effects. We assume that a large number of clinical trials have been

performed over the years for drugs A and C and that therefore there is only minor

Figure 1: Decision structure used in the illustrative case. Starting from the top, there is the decision goal (assessing value), that can be operationalized with four criteria. The relative importance of the criteria is indicated by the criterion weights and the plus or minus indicates if the criterion is to be maximized or minimized. The performance of the four decision alternatives at the bottom on the criteria is determined with performance scores. Note that for clarity only the arrows showing the performance scores for drug A are shown.

uncertainty surrounding their clinical performance. We assume there is a new drug “C” that is potentially much more effective than the conventional drugs. However, we assume that due to its novelty only a small number of patients have been enrolled in clinical trials. This means there still is considerable uncertainty regarding its actual clinical performance. It is assumed that placebo provides almost no effectiveness and that it is associated with very little adverse events. An overview of the datasets for preferences and clinical performances is presented in Table 1.

Criterion	Criterion weights (SD)		Performance (events / n)			
	Patients (n=100)	Experts (n=5)	Drug A	Drug B	Drug C	Placebo
Probability of response (+)	0.46 (0.04)	0.01 (0.03)	5500 / 7000	6000 / 7000	84 / 100	250 / 1000
Probability of remission (+)	0.19 (0.02)	0.69 (0.07)	6000 / 7000	5000 / 7000	84 / 100	250 / 1000
Probability of adverse event (-)	0.14 (0.03)	0.13 (0.10)	300 / 7000	300 / 7000	1 / 100	5 / 1000
Probability of severe adverse event (-)	0.21 (0.02)	0.08 (0.08)	30 / 7000	30 / 7000	0 / 100	50 / 1000

Table 1: Hypothetical dataset used in the case study. Events=number of patients in trials that experience the event. N=total sample size of the trials. The + sign indicates that a criterion is to be maximized. Similarly, a – sign indicates that a criterion is to be minimized.

Case methods

In each simulation run t , criterion weights w_{kt} are obtained by using a bootstrap resampling method. This means that for each run a bootstrap sample of 100 cases of the weight dataset was drawn with replacement. Because the clinical performances of drugs are proportions, performance samples θ_{ikt} are assumed to be distributed with a Beta distribution [21]. After sampling from the Beta distributions, $v_{kt}(\theta_{ikt})$ and V_{it} are calculated using Formula's 1 and 2.

In total, $T = 10,000$ Monte Carlo simulations are performed. The 95% confidence interval for each treatment's value was estimated with the 2.5%th and 97.5%th quantiles of its V_{it} from the simulation output. The model was programmed in R [24]. Several scenario analyses will be performed. First of all, a model that uses only the mean criterion weights and mean performance scores will be run. Then, the impact of uncertainty will be explored by running separate Monte Carlo simulations with

1. only uncertainty in criterion weights (that is, fixing the performances at their mean values while varying the weights as in the base case),
2. only uncertainty in performance scores (that is, fixing the weights at their mean values while varying the performances as in the base case), or
3. uniform probability distributions for criterion weights (keeping the sum of weights constant at one, and varying the performance scores as in the base case).

Case results

Patient and expert valuations of drugs

When using a deterministic model, that is, setting both criterion weights and performance scores to their mean values, the overall scores for drug A, drug B, drug C and placebo are 0.51, 0.51, 0.54, and 0.15, respectively for patients. For experts, the overall scores for drug A, drug B, drug C and placebo would be 0.67, 0.57, 0.67, and 0.19, respectively. This suggests that drug C has the highest value for patients and drugs A and C seem the most valuable treatment according to experts. Although this is already an insightful result, we cannot assess the confidence of these valuation statements. Taking into account uncertainty as described in the methods section gives us more insight into the treatment valuation by patients and experts (Figure 2). Note the more spread out probability density for the value of drug C which indicates that its value is more uncertain than that of drugs A and B. Drug C seems still to be the most valuable treatment to patients with a probability of being ranked first of 73%. Placebo still has the last rank, as would be expected. There is considerable uncertainty in the treatment values: there is a 27% probability that drug C turns out to not be the most valuable to patients. Furthermore, there is considerable decision uncertainty as to the second most valuable drug ($r_{2A} = 37\%$ and $r_{2B} = 47\%$). The clinical experts' results incorporating uncertainty show that drugs A and C that both have a score of 0.67.

The first rank probabilities for drug A and drug C are 51% and 49%, respectively. This means there is clinical equipoise between drugs A and C according to experts. Drug B is ranked third in all simulations with a score of 0.57 and placebo is again ranked last in all simulations with a score of 0.19. The impact of patient preferences as opposed to clinical experts is thus that while patients seem certain that drug C has the highest value, experts consider drugs A and C to be equally valuable.

Impact of uncertainty

When uncertainty in either patient-assigned criterion weights or performance scores is ignored, the point estimates for all four drugs remain similar. However, the confidence intervals of the drugs become smaller. This can be seen also in the ranking probabilities, which are higher for each rank. The rank reversal probabilities for first rank decreases to 18% when uncertainty in performances is not taken into account and increases to 41% when uncertainty in criterion weights is not taken into account. This means that here performance uncertainty seems to have a larger impact than preference uncertainty on the confidence with which the most valuable drug is chosen. Finally, using a uniform distribution for criterion weights induced a large variation in drug scores and consequently,

a high rank reversal probability for the first rank (61%). This large variation in scores is logical because the criterion weights vary between 0 and 1, whereas in the other scenarios there is much less variation in sampled weights (Table 1).

Scenario	Parameter	Drug A	Drug B	Drug C	Placebo
<i>Patients</i>	Score (95% CI)	0.51 (0.48 to 0.54)	0.51 (0.48 to 0.54)	0.54 (0.49 to 0.58)	0.15 (0.13 to 0.17)
	P(Rank=1)	10%	17%	73%	0%
	P(Rank=2)	37%	47%	16%	0%
	P(Rank=3)	53%	36%	11%	0%
	P(Rank=4)	0%	0%	0%	100%
<i>Experts</i>	Score (95% CI)	0.67 (0.65 to 0.68)	0.57 (0.56 to 0.59)	0.67 (0.61 to 0.71)	0.19 (0.17 to 0.21)
	P(Rank=1)	51%	0%	49%	0%
	P(Rank=2)	49%	0%	51%	0%
	P(Rank=3)	0%	100%	0%	0%
	P(Rank=4)	0%	0%	0%	100%
<i>Patients, only preference uncertainty</i>	Score (95% CI)	0.52 (0.50 to 0.55)	0.53 (0.50 to 0.55)	0.55 (0.52 to 0.58)	0.15 (0.14 to 0.16)
	P(Rank=1)	6%	12%	82%	0%
	P(Rank=2)	36%	49%	14%	0%
	P(Rank=3)	58%	38%	4%	0%
	P(Rank=4)	0%	0%	0%	100%
<i>Patients, only performance uncertainty</i>	Score (95% CI)	0.54 (0.51 to 0.56)	0.54 (0.51 to 0.57)	0.55 (0.50 to 0.60)	0.17 (0.16 to 0.19)
	P(Rank=1)	15%	26%	59%	0%
	P(Rank=2)	37%	43%	20%	0%
	P(Rank=3)	48%	31%	20%	0%
	P(Rank=4)	0%	0%	0%	100%
<i>Uniform distributions for criterion weights</i>	Score (95% CI)	0.40 (0.12 to 0.68)	0.38 (0.11 to 0.66)	0.42 (0.14 to 0.70)	0.11 (0.02 to 0.20)
	P(Rank=1)	33%	28%	39%	0%
	P(Rank=2)	34%	33%	33%	0%
	P(Rank=3)	31%	37%	26%	7%
	P(Rank=4)	2%	3%	2%	93%

Table 2: model outcomes: overall scores and rank probabilities.

Figure 2: Probability density estimation plot (Gaussian kernel estimation using the *density* function in R) of the model results for when patient preferences are used. Red=Drug A, green=Drug B, Blue=drug C and purple=placebo.

Impact of uncertainty

When uncertainty in either patient-assigned criterion weights or performance scores is ignored (that is, set to their mean values), the point estimates for all four drugs remain similar. However, the confidence intervals of the drugs become smaller. This can be seen also in the ranking probabilities, which are higher for each rank. The rank reversal probabilities for first rank decreases to 18% when uncertainty in performances is not taken into account and increases to 41% when uncertainty in criterion weights is not taken into account. This means that in the case performance uncertainty seems to have a larger impact than preference uncertainty on the confidence with which the most valuable drug is chosen. Finally, using a uniform distribution for criterion weights induced a very large variation in drug scores and consequently, a high rank reversal probability for the first rank (61%). This large variation in scores is logical because the criterion weights vary between 0 and 1, whereas in the other scenarios there is much less variation in sampled weights (Table 1).

Discussion

In this paper we have demonstrated a probabilistic multi-criteria approach to determine the patient-weighted value of treatments. The MCDA model developed for this purpose takes into account the parameter uncertainty surrounding both the elicited preferences and clinical trial data. The model was illustrated using a hypothetical case on three

Figure 3: Overview of the overall mean value of treatments for the patients (base case) (a), for the clinical experts (b), as well as for the scenario analyses investigating the impact of uncertainty in criterion weights (c), performance scores (d) and the impact of using non-informative uniform distributions for criterion weights (e). The error bars indicate the 95% confidence intervals. Pts=Patients, Plc=Placebo.

antidepressant treatments and placebo. In the case the patient-weighted treatment values are considerably different from the expert-weighted values. Furthermore, the rank order of treatments is still uncertain for patients and experts (as reflected in the rank reversal probabilities). Scenario analyses showed that in this case decision uncertainty seems to depend more on uncertainty in clinical evidence than on uncertainty in patient preferences. Finally, adopting uniform criterion weight distributions lead to the most decision uncertainty, as reflected in the high probabilities of rank reversal.

Comparison to earlier work

Our MCDA model builds on and combines characteristics from earlier approaches. First of all, it draws on the MCDA approach for the model structure and value functions [3,7]. Secondly, preference data from stated preference methods can be included in the model, allowing the incorporation of uncertainty around patient preferences. Thirdly, this uncertainty was combined with uncertainty around clinical performance estimates using Monte Carlo simulation methods. Although there have been other methods to combine patient preferences and clinical trial data in the context of healthcare policy, these are mainly limited by not practically taking into account multiple (concurrent) events and/or uncertainty around preferences [25–29]. Stochastic multi-criteria acceptability analysis (SMAA) also combines preference data with clinical trial data, but a non-informative (uniform) distribution or a single rank order of criteria is used for preferences [19,30]. A

similar approach is adopted by Caster *et al* who include a rank order of criteria importances based on qualitative information on utilities [31]. Although both SMAA and the method by Caster *et al* can include information about patient preferences, only including rank orders of criteria would preclude decision makers from considering the rich information on patient preferences yielded by stated preference studies.

Applicability and advantages of the model

As this was a demonstration and not an empirical comparison of the model to other modes of decision making, strong claims about its advantages in practice cannot yet be made. However, we believe the presented approach may have several advantages for decision makers confronted with a treatment valuation task. First of all, the adopted MCDA approach can help decision makers to structure the available preference data and clinical evidence and can help them assess the impact of preferences on the overall value of treatments. The present study adds to this the explicit inclusion and combination of patient preferences and clinical evidence. Furthermore, to account for uncertainty in both preferences and clinical data, the flexible probabilistic approach is adopted. These two additions may give decision makers more insight into 1) the influence of patient preferences on treatment value, and 2) into the impact of uncertainty in both preferences and clinical data on the decision. A final advantage is that because of the explicit use of evidence and the use of visualizations, decision makers can use the model to communicate their decision (argumentation) to stakeholders. This can be especially true for communicating a decision to patients because patient preferences are explicitly used.

Our model considers mainly the evidence-based treatment valuation task, whereas a complete decision making process has much more steps. Therefore, real-world applications of our model would require it to be applied in the context of an overarching decision making framework that guides the decision making process from problem definition to final decision. One such framework is ProACT-URL, which structures the decision making process with the following phases: problem, objective, alternatives, consequences, trade-off, uncertainty, risk tolerance and linked decisions [32]. For an inclusion of the developed model in ProACT-URL, criterion weights should be elicited from patients in the process after the definition of the 'effects table' in the consequences phase. These could be combined with the clinical evidence according to the model developed in this study to help decision makers assess the benefit-risk balance from a patient's perspective in the trade-offs phase of ProACT-URL and to guide the discussion in the uncertainty phase that follows the 'trade-offs' phase. We argue that the inclusion of our proposed model into frameworks such as ProACT-URL would be most useful when it is judged by decision makers that the

decision to make is characterized by uncertain clinical evidence and/or uncertain patient preferences. Given the explicit use of elicited patient preference, decision makers seeking to apply our model should be aware of remaining normative issues regarding the use of elicited patient preferences in real world decision contexts. These are: whose preferences should be elicited, who should perform the preference elicitation study, and what stated preference method should be used.

Aspects of the decision making process that may change in real world policy decision contexts compared to our simple illustrative case, are that more patients could be involved and that more criteria (not all relating directly to the patient experience) may be considered relevant by the decision makers. Given previous experience with performing large patient preference studies [17,18,33] and experience with using MCDA to consider large amounts of criteria [3,34], it is reasonable to expect that our model can be extended to real world use. Furthermore, even if the real-life decision involves other criteria requiring other normative judgments outside the patient experience (such as societal willingness to pay), it is possible to construct an MCDA model that includes the preferences of multiple stakeholder groups. The relative weight of the preferences of these (and potentially other) stakeholder groups can then be weighed by the decision makers, who make the final decision after considering the outcomes of the evidence synthesis as facilitated by the MCDA model presented in this study. Finally, in the case of benefit-risk assessments decision makers may be reluctant to aggregate benefits and risks into one score [26]. In that case, decision makers could elect to model benefits and risks in separate MCDA models and use the results during the assessment of the benefit-risk assessment.

Limitations of the model and opportunities for further research

Our model has some limitations. First of all, a simple multi-attribute method with an additive value function that assumes independent criteria was used in this study. Although such methods allow for non-linear value functions, in this study the overall treatment value was assumed to scale linearly with the criterion weights. Since the lower and upper levels were 0% and 100% for all criteria in the case, this implied that criterion weights reflected the relative importance of events and non-linear preferences for probabilities were not incorporated. There are several categories of uncertainty in MCDA [22]. In this study, only parameter uncertainty was considered, while patient-specific preference variation and patient-specific variation in outcomes is increasingly becoming important in light of recent developments in personalized medicine [41]. A final and practical limitation is that the process of gathering relevant data on patient preference and clinical evidence, as well as building the model can be time-consuming. What MCDA methods to use in real-world

applications of the presented model should be the topic of future research. Aspects we believe are important include the type of patient preferences that are to be elicited (since these need to match the MCDA method [36]), the preferable type of clinical evidence and specific decision maker needs. It may be useful to look into experiences in other disciplines where there is a longer history of using MCDA to support decision makers (see e.g. [37–40]).

Conclusions

In conclusion, we have developed a novel approach to estimate the value of treatments from the patient's perspective using a probabilistic MCDA model. The model was illustrated with a case on antidepressants. The model can provide insight into the patient-weighted value of treatments and how this may differ from an expert's assessment. It also can provide insight into the impact of uncertainty that still surrounds the value of treatments. Future work will need to address patient-specific variation and the feasibility of the modeling approach in practical applications (specifically in existing regulatory decision making frameworks).

List of abbreviations

CI: confidence interval, MCDA: multi criteria decision analysis, MACBETH: measuring attractiveness by a categorical based evaluation technique, SMAA: stochastic multicriteria acceptability analysis, ProACT-URL: problem, objective, alternatives, consequences, trade-off, uncertainty, risk tolerance and linked decisions.

References

- [1] Diaby V, Campbell K, Goeree R. Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis. *Oper Res Heal Care* 2013;2:20–4.
- [2] Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the Value of Healthcare Interventions Using Multi-Criteria Decision Analysis: A Review of the Literature. *Pharmacoeconomics* 2014;32:1–21.
- [3] Belton V, Stewart TJ. Multiple criteria decision analysis : An integrated approach. 2nd ed. Dordrecht: Kluwer Academic Publishers; 2002.
- [4] Doggson J, Spackman M, Pearman A, Phillips L. Multi-criteria analysis : A manual. London: 2009.
- [5] Hughes D, Waddingham E, Mt-isa S, Goginsky A, Chan E, Downey G, et al. Recommendations for the methodology and visualisation techniques to be used in the assessment of benefit and risk of medicines. London: 2013.
- [6] Ishizaka A, Nemery P. Multi-criteria decision analysis: methods and software. 1st ed. John Wiley & Sons Ltd; 2013.
- [7] Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines — part 1 : the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiol Drug Saf* 2007;16:S2–15.

- [8] Felli JC, Noel RA, Cavazzoni PA. A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. *Med Decis Mak* 2009;29:104–15.
- [9] Hummel JM, Bridges JFP, IJzerman MJ. Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. *Patient* 2014;7:129–40.
- [10] Tony M, Wagner M, Khoury H, Rindress D, Papastavros T, Oh P, et al. Bridging health technology assessment (HTA) with multicriteria decision analyses (MCDA): field testing of the EVIDEM framework for coverage decisions by a public payer in Canada. *BMC Health Serv Res* 2011;11:329.
- [11] Diaby V, Goeree R. How to use multi-criteria decision analysis methods for reimbursement decision-making in healthcare: a step-by-step guide. *Expert Rev Pharmacoecon Outcomes Res* 2014;14:81–99.
- [12] Goetghebeur MM, Wagner M, Khoury H, Levitt RJ, Erickson LJ, Rindress D. Bridging health technology assessment (HTA) and efficient health care decision making with multicriteria decision analysis (MCDA): applying the EVIDEM framework to medicines appraisal. *Med Decis Mak* 2012;32:376–88.
- [13] Van Til JA, IJzerman MJ. Why Should Regulators Consider Using Patient Preferences in Benefit-risk Assessment? *Pharmacoeconomics* 2013;10–3.
- [14] Facey K, Boivin A, Gracia J, Hansen HP, Lo Scalzo A, Mossman J, et al. Patients' perspectives in health technology assessment: a route to robust evidence and fair deliberation. *Int J Technol Assess Health Care* 2010;26:334–40.
- [15] Bridges JFP, Jones C. Patient-based health technology assessment: a vision of the future. *Int J Technol Assess Health Care* 2007;23:30–5.
- [16] MDIC PCBR project group members. A framework for incorporating information of patient preferences regarding benefit and risk into regulatory assessments of new medical technology. Medical Device Innovation Consortium: 2015.
- [17] Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl Health Econ Health Policy* 2013;11:319–29.
- [18] Weernink MGM, Janus SIM, van Til J a., Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceut Med* 2014.
- [19] Van Valkenhoef G, Tervonen T, Zhao J, de Brock B, Hillege HL, Postmus D. Multicriteria benefit-risk assessment using network meta-analysis. *J Clin Epidemiol* 2012;65:394–403.
- [20] Briggs AH, Weinstein MC, Fenwick E a L, Karnon J, Sculpher MJ, Paltiel a D. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Heal* 2012;15:835–42.
- [21] Claxton K. Briefing paper for methods review workshop on exploring uncertainty. York: 2007.
- [22] Broekhuizen H, Groothuis-Oudshoorn C, van Til J, Hummel M, IJzerman M. A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics* 2015;33:445–55.
- [23] Hummel JM, Volz F, van Manen JG, Danner M, Dintsios CM, IJzerman MJ, et al. Using the analytic hierarchy process to elicit patient preferences. *Patient* 2012;5:1–13.
- [24] R Development Core Team. R: A Language and Environment for Statistical Computing 2012.
- [25] Holden W. Benefit-risk analysis: A brief review and proposed quantitative approaches. *Drug Saf* 2003;26:853–62.
- [26] Shaffer ML, Watterberg KL. Joint distribution approaches to simultaneously quantifying benefit and risk. *BMC Med Res Methodol* 2006;6.

- [27] Lynd LD, Najafzadeh M, Colley L, Byrne MF, Willan AR, Sculpher MJ, et al. Using the incremental net benefit framework for quantitative benefit-risk analysis in regulatory decision-making - a case study of alosetron in irritable bowel syndrome. *Value Heal* 2009;13:1–7.
- [28] Lynd LD, Marra CA, Najafzadeh M, Sadatsafavi M. A quantitative evaluation of the regulatory assessment of the benefits and risks of rofecoxib relative to naproxen: an application of the incremental net-benefit framework. *Pharmacoepidemiol Drug Saf* 2010;19:1172–80.
- [29] Wen S, Zhang L, Yang B. Two approaches to incorporate clinical data uncertainty into multiple criteria decision analysis for benefit-risk assessment of medicinal products. *Value Heal* 2014;17:619–28.
- [30] Tervonen T, van Valkenhoef G, Buskens E, Hillege HL, Postmus D. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Stat Med* 2011;30:1419–28.
- [31] Caster O, Norén GN, Ekenberg L, Edwards IR. Quantitative Benefit-Risk Assessment Using Only Qualitative Information on Utilities. *Med Decis Mak* 2012;32:E1–15.
- [32] Zafiroopoulos N, Phillips L, Pignatti F, Luria X. Evaluating benefit-risk: An agency perspective. *Regul Rapp* 2012;9:5–8.
- [33] Groothuis-Oudshoorn CG, Fermont JM, van Til J a, Ijzerman MJ. Public stated preferences and predicted uptake for genome-based colorectal cancer screening. *BMC Med Inform Decis Mak* 2014;14:18.
- [34] EMA. Work package 2 report: Applicability of current tools and processes for regulatory benefit-risk 2011;44:1–33.
- [35] Bojke L, Claxton K, Sculpher M, Palmer S. Characterizing structural uncertainty in decision analytic models: a review and application of methods. *Value Heal* 2009;12:739–49.
- [36] Choo EU, Schoner B, Wedley WC. Interpretation of criteria weights in multicriteria decision making. *Comput Ind Eng* 1999;37:527–41.
- [37] Huang IB, Keisler J, Linkov I. Multi-criteria decision analysis in environmental sciences : Ten years of applications and trends. *Sci Total Environ* 2011;409:3578–94.
- [38] Wang J-J, Jing Y-Y, Zhang C-F, Zhao J-H. Review on multi-criteria decision analysis aid in sustainable energy decision-making. *Renew Sustain Energy Rev* 2009;13:2263–78.
- [39] Steuer RE, Na P. Multiple criteria decision making combined with finance: A categorized bibliographic study. *Eur J Oper Res* 2003;150:496–515.
- [40] Ho W, Xu X, Dey PK. Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. *Eur J Oper Res* 2010;202:16–24.
- [41] Rogowski W, Payne K, Schnell-Inderst P, Manca A, Rochau U, Jahn B, et al. Concepts of “personalization” in personalized medicine: implications for economic evaluation. *Pharmacoeconomics* 2015;33:49–59.

Chapter 4

Weighing clinical evidence using patient preferences: an
application of probabilistic multi-criteria decision
analysis

Henk Broekhuizen

Maarten J. IJzerman

A. Brett Hauber

Catharina G.M. Groothuis-Oudshoorn

Published as: Broekhuizen H, IJzerman MJ, Hauber AB, Groothuis-Oudshoorn CGM.
Weighing clinical evidence using patient preferences: an application of probabilistic multi-
criteria decision analysis. *Pharmacoeconomics* 2017;35:259–69.

Abstract

The need for patient engagement has been recognized by regulatory agencies, but there is no consensus about how to operationalize this. One approach is the formal elicitation and use of patient preferences for weighing clinical outcomes. The aim of this study was to demonstrate how patient preferences can be used to weigh clinical outcomes when both preferences and clinical outcomes are uncertain by applying a probabilistic value-based multi-criteria decision analysis (MCDA) method. Probability distributions were used to model random variation in and parameter uncertainty in preferences, and parameter uncertainty in clinical outcomes. The posterior value distributions and rank probabilities for each treatment were obtained using Monte Carlo simulations. The probability of achieving the first rank is the probability that a treatment represents the highest value to patients. We illustrated our methodology for a simplified case on six HIV treatments. Preferences were modeled with normal distributions and clinical outcomes were modeled with beta distributions. The treatment value distributions showed the rank order of treatments according to patients and illustrate the remaining decision uncertainty. This study demonstrated how patient preference data can be used to weigh clinical evidence using MCDA. The model takes into account uncertainty in preferences and clinical outcomes. The model can support decision makers during the aggregation step of the MCDA process and provides a first step toward preference-based personalized medicine, yet requires further testing regarding its appropriate use in real-world settings.

Introduction

Assessing the relative effectiveness of treatments is an essential component in regulatory and reimbursement decisions and is substantiated by strong methodology and clinical evidence development guidelines. However, interpretation of clinical evidence is a largely subjective and opinion-based process and such judgments are rarely formally included in decision making processes [1]. Nevertheless, subjective value judgments of stakeholders (in particular those of patients [1–4]), are an essential part of what finally determines treatment value [5]. For instance, in decisions considering two or more equally effective drugs, the (subjective) relative severity of the associated adverse events may dominate the final decision. Currently, subjective value judgments such as the relative severity of adverse events are mostly considered implicitly in healthcare policy decisions. Patient engagement is increasingly promoted [5–7], and several mechanisms of patient engagement have been used, including patient panels and patient reported outcomes. Whereas decision makers can use such implicit viewpoints, these can still be inaccurate or biased [8,9]. A potentially more representative and transparent approach is the use of results from survey-based stated preference studies [10]. These studies yield numeric estimates of the relative importance that respondents place on attributes of medical services such as clinical outcomes or other treatment characteristics [1,2].

Although elicited preferences can be used as a piece of information in the deliberative process of assessing the relative treatment value, a more formal approach would actually use such preference to weigh clinical endpoints and, thus, prioritize interventions by explicitly mapping the benefits and risks. One method that could be used to structure and analyze decision problems and to explicitly include patient preferences is multi-criteria decision analysis (MCDA). The application of MCDA allows decision makers to structure their decision problem and work on it in a transparent and consistent way [11,12]. Importantly, MCDA is flexible in that it allows multiple stakeholder groups, including patients, to assign preference weights to clinical outcomes (which are referred to as criteria in MCDA). An MCDA process typically comprises several steps, including the (1) definition of the decision problem and alternatives, (2) identification of the decision criteria, (3) weighting of the criteria, (4) identification of the performance of the alternatives, (5) scoring the performance, (6) aggregating the results and dealing with uncertainty, and (7) reporting the results [13].

The use of MCDA in healthcare has grown over the last few years [14,15] which has led to a recent ISPOR emerging good practices taskforce to publish guidance on the basic concepts and implementation of MCDA in healthcare [13,16].

The most commonly used method for aggregating the criterion weights and performance scores in MCDA is the linear additive value function. By combining the criterion weights derived from a group of patients and the actual clinical performance scores of the therapeutic alternatives, one can estimate patient-weighted treatment values [17]. This is a relatively straightforward approach but it only reflects the mean clinical treatment performance for an average patient and uncertainty in neither the criteria weights nor performance scores is evaluated. There are, however, several solutions to include uncertainty in the MCDA model. For instance, Wen *et al.* and Chim *et al.* use probability distributions for clinical evidence and point estimates for preferences [18,19]. Kaltoft *et al.* estimate preference uncertainty by defining preference subgroups and coupled these to point estimates for clinical performance [20]. Lynd *et al.* used point estimates from a discrete choice study to assign utilities to events in a (probabilistic) discrete event simulation [21]. In two other methods, probability distributions for clinical evidence is combined with uniform distributions for criterion weights as a non-informative prior [22,23]. One common problem of all these methods is that they do not simultaneously combine parameter uncertainty and random variation in both patient preferences and clinical evidence. Here, parameter uncertainty is uncertainty around an estimated quantity (such as a group mean) which can be reduced with more measurements [24,25]. Random preference variation is the systematic variation of preferences across the population. This variation can, by definition, not be reduced by repeated measurements but can only be better characterized [24,25]. The importance of including preference variation is recently recognized in the development of personalized medicine based on (variation in) preferences [26,27].

The aim of this study was to demonstrate an application of probabilistic MCDA that would allow a joint analysis of patient preferences and clinical evidence for treatments, taking into account uncertainty in both preferences and clinical evidence. The proposed model can be applied during the aggregation step of a value-based MCDA and is able to handle three sources of uncertainty: random preference variation, parameter uncertainty in preferences, and parameter uncertainty in clinical evidence. The model is designed to yield value distributions that enable explicit probabilistic statements about which treatments are preferred. The model is illustrated using a simplified case study of highly active anti-retroviral treatments for HIV patients.

A proposed MCDA method for capturing the value of treatments under uncertainty

Building the value function and defining sources of uncertainty

We adopted a value-based MCDA method. The value of a particular treatment i , denoted V_i , $i = 1, \dots, n$ was assumed to be a linear additive function of K criteria:

$$V_i = \sum_{k=1}^K \beta_k X_{ki}, \quad (1)$$

where β_k denotes the preference weight of criterion k , and X_{ki} is the performance of treatment i on criterion k . We assumed that patients prefer treatments with a higher value to treatments with a lower value, that preference weights and the clinical performances were not correlated, and that the clinical performances were measured on an interval scale.

To introduce random preference variation, suppose each patient q has his/her own preference weight β_{kq} for each criterion k :

$$V_{iq} = \sum_{k=1}^K \beta_{kq} X_{ki}. \quad (2)$$

The term β_{kq} in this equation is composed of population mean preference weight β_k and a patient-specific random effect θ_{kq} :

$$\beta_{kq} = \beta_k + \theta_{kq}, \quad (3)$$

with

$$\theta_{kq} \sim N(0, \sigma_k^2). \quad (4)$$

Parameter estimators

There are different methods for obtaining preference weights, such as swing weighting. However, in this study the required parameter estimators $\hat{\beta}_k$ and $\hat{\sigma}_k^2$ were obtained from analyzing patient-level data from stated preference

Figure 1: Overview of the Monte- Carlo simulation method used in the model with $i = 1, \dots, n$ the treatment, and $t = 1, \dots, T$ the Monte Carlo simulation run

studies. Parameter uncertainty in these estimators across all criteria is reflected in the covariance matrices $\Sigma_{\hat{\beta}}$ and $\Sigma_{\hat{\sigma}^2}$. The estimator for clinical performance X_{ki} is denoted $\hat{x}_{ki}$ and can be obtained from clinical trial reports.

Sampling framework

We assumed that the vectors $\hat{\beta} = (\hat{\beta}_1 \dots \hat{\beta}_K)$ and $\hat{\sigma} = (\hat{\sigma}_1 \dots \hat{\sigma}_K)$ were both distributed according to multivariate normal distributions. We denoted the value distribution of treatment i across the population with ψ_i , assuming that the individual values V_{iq} were identically distributed for patients q . In a probabilistic model all uncertain parameters are varied at the same time, and this implies that the probability distribution ψ_i is a complex and analytically challenging combination of the distributions for $\hat{\beta}_k$, $\hat{\sigma}_k$ and $\hat{x}_{ki}$. We therefore approximated ψ_i with Monte Carlo simulations (Figure 1). The Monte Carlo

simulations were programmed as follows. In each simulation run t , we first sampled a population mean preference weight $\boldsymbol{\beta}_t = (\beta_{1t} \dots \beta_{Kt})$ from $MVN(\hat{\boldsymbol{\beta}}, \Sigma_{\hat{\boldsymbol{\beta}}})$ and a standard deviation $\boldsymbol{\sigma}_t = (\sigma_{1t} \dots \sigma_{Kt})$ from $MVN(\hat{\boldsymbol{\sigma}}, \Sigma_{\hat{\boldsymbol{\sigma}}})$. From that we obtained a respondent q_t using Formulas 3 and 4:

$$\boldsymbol{\beta}_{q_t} \sim MVN(\boldsymbol{\beta}_t, \Sigma_{\sigma_t^2}), \quad (5)$$

with $\Sigma_{\sigma_t^2}$ a diagonal covariance matrix with $\boldsymbol{\sigma}_t^2 = (\sigma_{1t}^2, \dots, \sigma_{Kt}^2)$ on the diagonal. This means that in every simulation run t a hypothetical patient q_t with a particular vector of preferences $\boldsymbol{\beta}_{q_t} = (\beta_{1q_t} \dots \beta_{Kq_t})$ was obtained. The probability distribution of $\hat{x}_{ki}$ was denoted with F_{ki} and chosen for each criterion depending on what best modeling practices recommend for that type of clinical performance [24]. From F_{ki} we sampled X_{kit} in each simulation run. The preference sample was combined with samples from the probability distributions F_{ki} for each criterion k to calculate the value for each treatment i in the simulation run with

$$V_{iq_t} = \sum_{k=1}^K \beta_{kq_t} X_{kit}. \quad (6)$$

The Monte Carlo simulation process was repeated a large number of times T and was programmed in R [28].

Model outcomes

From the Monte Carlo simulations the mean value of each treatment in the population was obtained, calculated as the posterior mean, $V_i = \frac{\sum_{t=1}^T V_{iq_t}}{T}$. This metric can be interpreted as the mean perceived value of the treatment's clinical performance according to patients. Since we assumed that patients prefer treatments with a higher values to treatments with a lower values, the treatment value should only be interpreted relative to the value of other treatments. The degree to which the value distributions of two treatments overlap indicates how uncertain we are about selecting the treatment with the highest value. More concretely, to make probabilistic statements about the degree to which we are sure about which treatment has the highest value in each simulation t treatments were ranked from the first (highest value) rank to the last (lowest value) rank based on their respective v_{iq_t} . The rank achieved by treatment i in simulation run t was denoted with r_{it} . We denoted the rank of treatment i with R_i . The probability that this rank equals y was estimated with the percentage of simulations treatment i had rank y :

$$\frac{\sum_{t=1}^T 1_y(r_{it})}{T}, \quad (7)$$

$$\text{with } 1_y(r_{it}) = \begin{cases} 1, & \text{if } r_{it} = y \\ 0, & \text{otherwise} \end{cases}. \quad (8)$$

Since we assumed patients prefer treatments with a higher value to treatments with a lower value, the treatment with the highest estimated probability of being ranked first was considered the most preferred treatment. One minus the probability that the most preferred treatment is ranked first is the rank reversal probability for the first rank. The rank reversal probability for the first rank is used as a measure of decision uncertainty, because it is the probability that the most preferred treatment (according to our model) turned out not to be the most valuable treatment.

An illustration of the proposed approach on a simplified HIV case

To illustrate the proposed model, it is applied to the case of HIV treatments. The MCDA was designed according to the guidelines as proposed by the ISPOR taskforce [13]. Although the data used comes from real studies, the main purpose of this paper is to illustrate the modeling approach. The simulation results presented in this paper can therefore not be used to inform clinical HIV decisions. In the supplementary material of the published paper all used data (with references) as well as the R script can be found to replicate our results.

Identification of treatment alternatives and decision criteria

The example is about the comparison of the relative value of highly active antiretroviral therapies (HAART) for HIV-positive patients from a regulatory perspective. Treatments under consideration are those recommended by the National Institute of Health (NIH) for treatment-naïve patients [29]. A HAART consists of an active drug component and one of two backbones (Abacavir/Lamivudine (AL) or Tenofovir/Emtricitabine (TE)). The treatments included in the case study were four combination treatments (Dolutegravir+AL/TE¹, Efavirenz+AL, Raltegravir+TE, Atazanavir/Ritonavir+TE, Elvitegravir/Cobicistat+TE and Darunavir/Ritonavir+TE) plus two backbone-only treatments (AL and TE). For simplicity we will denote combination treatments only by their active drug component (e.g. Dolutegravir instead of Dolutegravir+AL/TE, etc) in the remainder of this paper.

¹ Either AL or TE

	Levels used in DCE [21]	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\sigma}_k$	$SE(\hat{\sigma}_k)$
<i>Virological failure prevented</i>	96.0%, 85.0%, 79.0%	-0.05	0.01	0.05	0.02
<i>Allergic reaction</i>	0.0%, 1.0%, 8.0%, 12.0%	-0.06	0.01	0.06	0.02
<i>Bone damage (treatable)</i>	0.0%, 1.0%, 5.0%, 10.0%	-0.01	0.02	0.06	0.06
<i>Kidney damage (not treatable)</i>	0.0%, 1.0%, 5.0%, 10.0%	-0.22	0.04	0.21	0.05

Table 1: Used preference data from [30]. All $\hat{\beta}$ are per percentage point probability of the event occurring in the next 52 weeks, i.e. the partial value of a 2% probability of allergic reaction in the next 52 weeks is -0.12. Note that the covariance $\Sigma_{\hat{\beta}}$ and $\Sigma_{\hat{\sigma}^2}$ are not presented here for brevity but can be found in the supplementary material. Both $\hat{\beta}_k$ and $\hat{\sigma}_k$ are assumed to be distributed with a multivariate normal distribution. DCE=discrete choice experiment

HAART	P(vir)	P(all)	P(bone)	P(kid)
<i>Dolutegravir+ TE/AL</i>	7.61% (6.10% to 9.25%)	0.24% (0.01% to 0.90%)	-	0.09% (0.00% to 0.34%)
<i>Atazanavir/Ritonavir + EL</i>	5.25% (3.94% to 6.77%)	1.72% (0.74% to 3.10%)	1.29% (0.48% to 2.5%)	0.43% (0.05% to 1.21%)
<i>Elvitegravir/Cobicistat + EL</i>	13.52% (10.17% to 17.34%)	-	2.02% (0.83% to 3.77%)	0.87% (0.18% to 2.12%)
<i>Efavirenz + AL</i>	13.67% (12.32% to 15.13%)	1.86% (1.21% to 2.66%)	2.12% (1.41% to 2.98%)	0.29% (0.11% to 0.56%)
<i>AL (backbone)</i>	16.3% (13.74% to 18.99%)	2.32% (1.01% to 4.18%)	4.57% (3.02% to 6.38%)	2.50% (1.47% to 3.75%)
<i>TE (backbone)</i>	20.41% (17.58% to 23.39%)	1.01% (0.28% to 2.19%)	2.19% (1.18% to 3.52%)	3.84% (2.56% to 5.36%)

Table 2: Used clinical evidence with 95% confidence intervals. For references to all included clinical trials, see the National Institute of Health guideline [29] and this manuscript's supplementary material. All performances were defined over a 52 week time horizon and assumed to be distributed with beta distributions. HAART=highly active anti-retroviral therapy, CI=confidence interval, AL=Abacavir/Lamivudine, TE=Tenofovir/Emtricitabine, P(vir)=probability of virological failure, P(all)=probability of allergic reaction, p(bone)=probability of bone damage, p(kid)=probability of kidney damage.

In the MCDA model we used the same criteria as identified in an earlier preference study [30], namely: probability of virologic failure (p_{viro}), probability of allergic reaction (p_{all}), probability of bone damage (p_{bone}) and probability of kidney damage (p_{kid}) [30]. All probabilities were all defined over a 52 week time horizon.

Measuring performance of the alternatives

The clinical performance of the treatments on the included criteria was obtained from clinical trial reports (for an overview of the included trials see the NIH guideline [29]). All criteria were represented as probabilities. For each treatment we therefore retrieved from the clinical trial reports the number of events e_k^+ and number of non-events e_k^- per criterion k . The performance estimates for each criterion k were then calculated with $\frac{e_k^+}{e_k^+ + e_k^-}$. For studies that reported time horizons other than 52 weeks, we used linear extrapolation to obtain performance estimates for a 52 week time horizon. We used a linear partial value function for each criterion. No rescaling of the performances was needed for the partial value functions since probabilities are already naturally constrained between 0 and 1. For criteria measured on other scales, such a rescaling to the $[0,1]$ range using a predefined lower and upper level would be required [31].

Plasma HIV RNA of more than 50 copies per mL 52 weeks after treatment start was considered a virologic failure event. Reported incidence of rash was used as a surrogate measure for the TE-induced allergic reaction event. Reported fractures in the clinical trials were considered to constitute bone damage events. For TE this was not reported; therefore decreases in bone mineral density of more than 6% were used as a surrogate endpoint. Because of the diagnostic and treatment options available, we limited our model to include only treatable bone damage [29,32]. Reported cases of renal failure were considered to be kidney damage events, and we therefore limited our model to include only non-treatable kidney damage [33].

Input for criteria weights: results from a previously published DCE experiment

The present methodological study did not itself elicit criteria weights, but used the results from a previous stated preference study to inform the preference weights. In that study by Hauber *et al*, 147 treatment-naïve HIV-positive African Americans gave their preferences for four criteria relevant for HIV treatment [30]. The study employed a discrete choice study design with 24 choice tasks. In the choice tasks the criteria were defined as the probability of the event in the next 52 weeks. An overview of the preference data is given in Table 1.

Aggregating scores and performance

In the base-case analysis we assumed an additive value function. The value of treatment i for patient q (Formula 2) could therefore be specified as:

$$V_{iq} = p_{viro,i} \beta_{viro,q} + p_{all,i} \beta_{all,q} + p_{bone,i} \beta_{bone,q} + p_{kid,i} \beta_{kid,q}. \quad (10)$$

The mean population preferences and the random preference variations were assumed to be normally distributed. The parameter estimates for these distributions were obtained from a (mixed logit [34]) analysis of the used preference study (Table 1). The performances of each treatment are probabilities and were therefore modeled with beta distributions [24]. Beta distributions require two input parameters (α_1 and α_2) and these were estimated from the clinical trials cited in the National Institute of Health guideline. A complete overview of the clinical trials used in this study can be found in the online supplementary material. We used the number of reported events e_k^+ as α_1 and the number of non-events e_k^- as α_2 . In total, 100,000 Monte Carlo simulations were performed.

Handling uncertainty: scenario analyses

A scenario analysis was employed to illustrate the effect of the different sources of uncertainty on the treatment value distributions. The base case scenario described in the previous paragraphs considered all types of uncertainty simultaneously, i.e. parameter uncertainty in preferences, random preference variation, and parameter uncertainty in clinical evidence.

Each of the other scenarios tested for one specific source of uncertainty. This means that each scenario included one source of uncertainty while the other two sources were assumed to have no uncertainty and were thus fixed at a particular value. In the first scenario, only parameter uncertainty in preferences (as parameterized by the covariance matrix Σ_{β}) was taken into account. Therefore θ_{kq} was set to zero for all criteria and individuals, and all clinical performances were set to the mean clinical performance of treatments as found in Table 2. In the second scenario, only random preference variation was included. This means that the Σ_{β} was set to be a zero matrix, all clinical performances were set to the mean clinical performance of treatments as found in Table 2, and θ_{kq} was distributed as in the base case. In the third scenario, only parameter uncertainty in clinical performance was included. This means that θ_{kq} was set to zero for all criteria and individuals, Σ_{β} was set to be a zero matrix, and performances X_{ki} were distributed with F_{ki} as in the base case.

Examination and discussion of findings

Outcomes of case study

The model outcome was a value distribution for each included HAART (Figure 2). In the base case, Dolutegravir has the highest patient-weighted estimated mean value of -0.39 with an empirical 95% confidence interval running from -1.25 to 0.48. The backbone-only treatments AL and TE had the lowest mean values (-1.49 and -1.86, respectively). In all scenario analyses, mean treatment values were similar to the base case results (Figure 2, Table 3) and the most likely rank order of treatments did not change. The width of the confidence intervals did vary between the base case and the scenarios. When only random preference variation was considered in the second scenario, the confidence intervals were only slightly narrower than those of the base case. In the two other scenarios that considered parameter uncertainty the confidence intervals were substantially narrower.

In 49.1% of the simulations, Dolutegravir was ranked first (Table 4). This implies that in 50.9% of simulations, another treatment was preferred. Atazanavir/Ritonavir had the highest probability of being ranked second (40.0%) and Efavirenz had the highest probability of being ranked third (40.3.0%). The narrower value distributions in the scenarios is reflected in the rank probabilities. The rank probabilities in the scenario that considered only parameter uncertainty in preferences were higher: the probability of Dolutegravir attaining first rank and Atazanavir/Ritonavir attaining second rank were both more than 90%. This means that we were confident about the ranking of treatments for the first and second rank. Similarly, the first and second rank probabilities for Dolutegravir and Atazanavir/Ritonavir are more than 75% in the scenario that considers only parameter uncertainty in clinical performances. For the fourth until last rank, there is slightly more decision uncertainty in the scenario that considers parameter uncertainty in preferences than there is in the scenario that considers parameter uncertainty in clinical performances.

Implications of the modeling framework for personalized medicine

The main objective of this study was to develop a novel methodology. The HIV case was used to demonstrate the concepts and cannot be used for guidance on HIV treatment decisions. However, the modelling framework does allow an exploration of its usefulness for other applications: assessing the impact of uncertain parameters (clinical evidence and preferences) on decision uncertainty and personalizing treatment based on preferences.

Figure 2: Barplots of the regimen values with 95% confidence intervals across the four analysis scenarios. A=all types of uncertainty, B=only parameter uncertainty in preferences, C=only patient-specific preference variation, D=only parameter uncertainty in performances, *Purple* dolutegravir, *dark blue* elvitegravir/cobicistat, *light blue* atazanavir/ritonavir, *green* efavirenz, *yellow* abacavir/lamivudine, *red* tenofovir/emtricitabine

HAART regimen	Base case	Scenario analyses		
	All three types of uncertainty	Only parameter uncertainty in preferences	Only patient-specific preference variation	Only parameter uncertainty in performances
<i>Dolutegravir+TE/AL</i>	-0.39 (-1.25 to 0.48)	-0.38 (-0.57 to -0.19)	-0.39 (-1.24 to 0.47)	-0.38 (-0.48 to -0.30)
<i>Atazanavir/Ritonavir+EL</i>	-0.46 (-1.2 to 0.24)	-0.45 (-0.62 to -0.28)	-0.45 (-1.14 to 0.23)	-0.45 (-0.64 to -0.32)
<i>Elvitegravir/Cobicistat+EL</i>	-0.83 (-2.52 to 0.8)	-0.82 (-1.22 to -0.42)	-0.83 (-2.41 to 0.74)	-0.83 (-1.13 to -0.59)
<i>Efavirenz+AL</i>	-0.83 (-2.4 to 0.77)	-0.82 (-1.19 to -0.44)	-0.83 (-2.41 to 0.76)	-0.82 (-0.92 to -0.73)
<i>AL backbone</i>	-1.49 (-3.77 to 0.83)	-1.47 (-2.06 to -0.89)	-1.48 (-3.7 to 0.75)	-1.48 (-1.79 to -1.20)
<i>TE backbone</i>	-1.86 (-4.68 to 1.07)	-1.85 (-2.54 to -1.17)	-1.85 (-4.64 to 0.94)	-1.86 (-2.21 to -1.55)

Table 3: Values (with 95% confidence interval) for the included HAART regimens across the four analyses. HAART=highly active anti-retroviral therapy. AL=Abacavir/Lamivudine, TE=Tenofovir/Emtricitabine

The assessment of the impact of the uncertainty in model parameters on decision uncertainty is important for identifying the value of additional research: it is most worthwhile to further investigate model parameters where more information would most likely reduce decision uncertainty.

In the presented case, there is a clear difference between the impact of uncertainty types. When only parameter uncertainty in either preferences or performances was considered, one treatment was clearly the most valuable treatment. However, when random preference variation was considered, there was considerable overall decision uncertainty. The value of additional clinical research may therefore not be high but decision aids (that help patient think about their preferences, reducing their individual preference uncertainty) may be valuable.

This relates to the second application for which our work may be used: personalized medicine. Although most research in that field has focused on personalizing treatment based on clinically measurable patient characteristics such as genetic differences, there is an increasing interest in personalizing treatment based on patient preferences [26]. Our model could help decision makers to make the first steps toward such a personalization by showing to what extent (uncertainty in) patient preferences influences the choice of treatment.

A next step to formalize the assessment of the extent to which differences in preferences are relevant would be to calculate metrics such as the value of heterogeneity [35]. This metric shows the marginal population-wide value gained from having patients choose between more than one treatment. This value will be low if there is a clear most valuable treatment for all patients, but it will be high if there is clinical equipoise and/or much patient-specific preference variation. An important outstanding research issue here is how values derived from an MCDA can be contrasted with financial costs [36]. Further research could also be directed toward the integration of patient-specific clinical outcome measures; that is, introducing a patient-specific performance distribution that yields estimates of the performance of a specific treatment for a specific patient. For the HIV case this could, for example, be operationalized by estimating the treatability of kidney damage based on respondent characteristics such as level of kidney functioning at treatment start. Such a holistic view of patient-specific variation in both preferences and clinical outcomes would be a step toward combining the two current viewpoints on personalized medicine [26].

		Dolutegravir	Atazanavir/ Ritonavir	Elvitegravir/ Cobicistat	Efavirenz	AL backbone	TE backbone	
		+TE/AL	+EL	+EL	+AL			
Base case	All three types of uncertainty	1	49.07%	34.21%	4.95%	4.28%	1.92%	5.57%
		2	36.12%	39.98%	10.53%	7.54%	4.08%	1.75%
		3	5.56%	8.69%	40.25%	40.31%	3.27%	1.92%
		4	3.46%	8.47%	39.96%	39.00%	5.69%	3.42%
		5	4.60%	3.52%	3.76%	4.96%	69.52%	13.64%
		6	1.19%	5.13%	0.55%	3.91%	15.52%	73.70%
Scenario Analyses	1: Only parameter uncertainty in preferences	1	90.42%	9.58%	0.00%	0.00%	0.00%	0.00%
		2	9.58%	90.26%	0.14%	0.02%	0.00%	0.00%
		3	0.00%	0.11%	46.26%	53.63%	0.00%	0.00%
		4	0.00%	0.05%	53.60%	46.35%	0.00%	0.00%
		5	0.00%	0.00%	0.00%	0.00%	100.0%	0.00%
		6	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
	2: Only patient-specific preference variation	1	53.28%	32.25%	2.34%	5.32%	1.59%	5.22%
		2	33.77%	46.18%	8.50%	6.43%	3.93%	1.19%
		3	4.68%	5.81%	42.71%	42.69%	2.07%	2.04%
		4	2.76%	8.52%	44.4%	37.84%	3.69%	2.79%
		5	4.41%	2.46%	1.90%	3.49%	79.51%	8.23%
		6	1.10%	4.78%	0.15%	4.23%	9.21%	80.53%
	3: Only parameter uncertainty in performances	1	77.2%	22.79%	0.01%	0.00%	0.00%	0.00%
		2	22.8%	76.41%	0.68%	0.11%	0.00%	0.00%
		3	0.00%	0.74%	51.00%	48.26%	0.00%	0.00%
		4	0.00%	0.06%	48.19%	51.63%	0.12%	0.00%
		5	0.00%	0.00%	0.12%	0.00%	95.2%	4.68%
		6	0.00%	0.00%	0.00%	0.00%	4.68%	95.32%

Table 4: Ranking probabilities for all included regimens across the four analyses. AL=Abacavir/Lamivudine, TE=Tenofovir/Emtricitabine, TE/AL=either Tenofovir/Emtricitabine or Abacavir/Lamivudine

Strengths, weaknesses and a comparison to existing literature

The first strength of the current study is that it has demonstrated a methodological approach for combining preference data and clinical data into one value metric, which may contribute to the ongoing attempts to integrate patient preference research in health technology assessment and market approval. A second strength of the study is that the developed model allows for the simultaneous consideration of the impact of three sources of uncertainty, i.e. random preference variation and parameter uncertainty in both clinical and preference estimates.

The developed model uses results from stated preference studies to inform value judgements in policy decisions from one of the most important stakeholders in healthcare: the patient. By including stated preference studies in the model, the patient preferences are incorporated in an explicit, structured and representative manner. Many different types of preference elicitation methods exist. In this study we used results from a discrete choice study, but in theory our model is able to handle preference weights obtained from a wide range of preference elicitation methods as long as a value function can be constructed and probability distributions can be assigned to weights. Finally, the possibility of including preference studies with large sample sizes allows for the investigation of variation and uncertainty in preferences. The impact of these was investigated by assigning informative probability distributions, which sets our study apart from earlier studies that have used point estimates [18–21] or non-informative (uniform) distributions [22,23].

The decision to use a probabilistic approach in the present study follows from a recent review identifying five approaches to deal with uncertainty in MCDA [25]. The approach adopted in this study seems most advantageous for our aims for a number of reasons. It is the approach that is best suited for dealing with the preferences of a group of stakeholders and is most able to consider multiple uncertain parameters [25]. Another advantage is that it is possible to implement Monte Carlo simulations as a flexible method that can combine all types of parametric and non-parametric probability distributions [24]. For these reasons, decision makers could apply the method during the aggregation and uncertainty steps of the MCDA process described in the recent ISPOR taskforce report [13]. This would be especially advantageous in policy decisions where the patient perspective is considered explicitly and where various sources of uncertainty are relevant to consider.

Even though the model may be appropriate given our aims, a limitation of all studies that use results from stated preference research is that a person's stated preference may not be the same as their revealed preference [37]. A limitation specific to the illustrative case study was that treatment value is assumed to be linearly related to the clinical performance and that we could extrapolate linearly. The linearity specification could not be rejected in

the preference study but this could also have been due to the sample size [30]. A related limitation is that we assumed that we could extrapolate performance measures to conform to the 52 week time horizon used in the preference study. Furthermore, a limitation is that random preference variation was assumed to be normally distributed. Although this is practical and a commonly made assumption in patient preference research [38], in our study it resulted in a small percentage of Monte Carlo simulations having sign reversals for the preference weights, mainly for the virologic failure criterion. In a larger patient sample it may have been possible to estimate a more specific functional form for the value function [39]. Performance samples for virologic failure all fell outside of the range for which preferences were elicited, which may have biased the value estimates.

Conclusion

In an attempt to explore new approaches for increasing patient engagement in healthcare policy decisions, the current paper presents a probabilistic MCDA model in which treatment values were estimated by weighting clinical trial evidence with results from a patient preference study. The model outcomes were patient-weighted probability distributions of relative treatment value and the respective rank probabilities. The developed model was illustrated using a simplified case study. The adopted probabilistic approach integrates random preference variation and parameter uncertainty in patient preferences with parameter uncertainty in clinical evidence using a Monte Carlo simulation method. Further research about the use of the modelling approach for non-simplified cases and the match to decision maker needs is required.

References

1. Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl. Health Econ. Health Policy* 2013; 11:319–29.
2. Weernink MGM, Janus SIM, van Til J a., Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceut. Med.* 2014
3. van Til JA, IJzerman MJ. Why Should Regulators Consider Using Patient Preferences in Benefit-risk Assessment? *Pharmacoeconomics* 2013;10–3.
4. MDIC PCBR project group members. A framework for incorporating information of patient preferences regarding benefit and risk into regulatory assessments of new medical technology. *Medical Device Innovation Consortium*; 2015.
5. Eddy D. Anatomy of a decision. *JAMA*. 1990;263:441–3.

6. Hofmann B, Cleemput I, Bond K, Krones T, Droste S, Sacchini D, et al. Revealing and Acknowledging Value Judgments in Health Technology Assessment. *Int. J. Technol. Assess. Health Care* 2015;30:579–86.
7. EMA. Reflection paper on benefit-risk assessment methods in the context of the evaluation of marketing authorisation applications of medical products for human use: report of the CHMP working group on benefit-risk assessment methods. London; 2008. Available from: http://www.emea.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/01/WC500069634.pdf
8. de Bekker-Grob EW, Essink-Bot ML, Meerdling WJ, Koes BW, Steyerberg EW. Preferences of GPs and patients for preventive osteoporosis drug treatment: A discrete choice experiment. *Pharmacoeconomics* 2009;27:211–9
9. Johnson FR, Hauber AB, Özdemir S, Siegel CA, Hass S, Sands BE. Are gastroenterologists less tolerant of treatment risks than patients? Benefit-risk preferences in Crohn's disease management. *J. Manag. Care Pharm.* 2010;16:616–28.
10. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier D a, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value Heal.* 2013;16:3–13.
11. Keeney R, Raiffa H. Decisions with multiple objectives. Cambridge: Cambridge University Press; 1976.
12. Belton V, Stewart TJ. Multiple criteria decision analysis : An integrated approach. 2nd ed. Dordrecht: Kluwer Academic Publishers; 2002.
13. Marsh K, Iizerman M, Thokala P, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making —Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Heal.* 2016;19:1–13.
14. Diaby V, Campbell K, Goeree R. Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis. *Oper. Res. Heal. Care* 2013;2:20–4.
15. Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the Value of Healthcare Interventions Using Multi-Criteria Decision Analysis: A Review of the Literature. *Pharmacoeconomics* 2014;32:1–21.
16. Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making—An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Heal.* 2015;19:1–13.
17. Hummel JM, Volz F, van Manen JG, Danner M, Dintsios CM, Iizerman MJ, et al. Using the analytic hierarchy process to elicit patient preferences. *The Patient.* 2012;5:1–13.
18. Wen S, Zhang L, Yang B. Two approaches to incorporate clinical data uncertainty into multiple criteria decision analysis for benefit-risk assessment of medicinal products. *Value Heal.* 2014;17:619–28.
19. Chim L, Salkeld G, Stockler MR, Mileschkin L. Weighing up the benefits and harms of a new anti-cancer drug: a survey of Australian oncologists. *Intern. Med. J.* 2015;45:834–42.
20. Kaltoft M, Turner R, Nielsen J, Cunich M, Salkeld G, Dowie J. Addressing preference heterogeneity in public policy by combining Cluster Analysis and Multi-Criteria Decision Analysis. *Health Econ. Rev.* 2014;5:1–11.
21. Lynd LD, Najafzadeh M, Colley L, Byrne MF, Willan AR, Sculpher MJ, et al. Using the incremental net benefit framework for quantitative benefit-risk analysis in regulatory decision-making - a case study of alosetron in irritable bowel syndrome. *Value Heal.* 2009;13:1–7.
22. Tervonen T, van Valkenhoef G, Buskens E, Hillege HL, Postmus D. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Stat. Med.* 2014;30:1419–28.
23. Caster O, Norén GN, Ekenberg L, Edwards IR. Quantitative Benefit-Risk Assessment Using Only Qualitative Information on Utilities. *Med. Decis. Mak.* 2012;32:E1–15.

24. Briggs AH, Weinstein MC, Fenwick E a L, Karnon J, Sculpher MJ, Paltiel a D. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Heal.* 2012;15:835–42.
25. Broekhuizen H, Groothuis-Oudshoorn CGM, van Til JA, Hummel JM, IJzerman MJ. A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics.* 2015;33:445–55.
26. Rogowski W, Payne K, Schnell-Inderst P, Manca A, Rochau U, Jahn B, et al. Concepts of “personalization” in personalized medicine: implications for economic evaluation. *Pharmacoeconomics* 2015;33:49–59.
27. Ho MP, Gonzalez JM, Lerner HP, Neuland CY, Whang JM, McMurry-Heath M, et al. Incorporating patient-preference evidence into regulatory decision making. *Surg. Endosc.* 2015
28. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2015. Available from: <http://www.r-project.org>
29. Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents. [Internet]. Natl. Inst. Heal. 2014. Available from: <http://aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf>
30. Hauber AB, Mohamed AF, Watson ME, Johnson FR, Hernandez JE. Benefits, risk, and uncertainty: preferences of antiretroviral-naïve african americans for HIV treatments. *AIDS Patient Care STDs.* 2009;23:1–6.
31. Dodgson J, Spackman M, Pearman A, Phillips L. Multi-criteria analysis : A manual [Internet]. London; 2009. Available from: http://eprints.lse.ac.uk/12761/1/Multi-criteria_Analysis.pdf
32. Walker Harris V, Brown TT. Bone loss in the HIV-infected patient: evidence, clinical implications, and treatment strategies. *J. Infect. Dis.* 2012;205 Suppl:S391-8.
33. Molitoris B. Acute kidney injury. In: Goldman L, Schafer A, editors. *Cecil Med.* 24th ed. Philadelphia, PA: Saunders Elsevier; 2011.
34. McFadden D, Train K. Mixed MNL models for discrete response. *J. Appl. Econom.* 2000;15:447–70.
35. Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. *Med. Decis. Mak.* 2009;27:112–27.
36. Thokala P, Duenas A. Multiple Criteria Decision Analysis for Health Technology Assessment. *Value Heal.* 2012;15:1172–81.
37. Kjær T. A review of the discrete choice experiment - with emphasis on its application in health care. *Heal. Econ. Pap.* 2005;1–139.
38. Hauber A, Gonzalez JM, Groothuis-Oudshoorn C, Prior T, DA M, Cunningham C, et al. Statistical methods for the analysis of discrete-choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force. *Value Heal.* 2016;
39. Van Der Pol M, Currie G, Kromm S, Ryan M. Specification of the utility function in discrete choice experiments. *Value Heal.* 2014;17:297–301.

Chapter 5

A health economic perspective on
the future of lung cancer screening

Henk Broekhuizen

Sebastian Hinde

Maarten J. IJzerman

Andrea Manca

Abstract

Failure to significantly improve the lung cancer mortality rate in recent years is associated with the high proportion of late stage presentation. Screening programs using chest X-ray (CXR), and low-dose computed tomography (LDCT) have been proposed as potential means to identify the disease at an earlier (pre-symptomatic) stage. However, despite clinical potential and positive health economic literature, the uptake of lung cancer screening policies is low. The aim of this paper was to investigate this low uptake based on a review of recent literature. After a short exposition of the major clinical trials and cost-effectiveness studies, we argue that previous studies underestimate implementation barriers by failing to account for the large initial costs of LDCT screening infrastructure. Furthermore, evidence suggests targeted screening of high risk subgroups based on traditional definitions may only identify 27% of cancers. We conclude that these barriers may be best addressed by a two-tier screening policy for lung cancer, as is employed in other cancer screening initiatives. Two technologies eligible for such a role in lung cancer screening are breath analysis and biomarkers. Such first-tier technologies have the potential to increase screening participation while reducing the demand for LDCT screens, representing an ideal solution to the problem faced. We review the current literature concerning these technologies and conclude that significant developments are needed before these technologies can perform the required function, but that without them meaningful lung cancer screening programs will remain untenable.

Introduction and background

Lung cancer is the leading cause of cancer-related death worldwide [1]. Despite recent advances in treatment options, the average 5-year survival of lung cancer is still only 9.5% compared to 85%, 87%, and 59% for prostate, breast, and colorectal cancer, respectively [2]. Current NICE guidelines in England and Wales are aimed at early diagnosis, and ongoing NHS awareness campaign aim to incentivize persons experiencing lung cancer-like symptoms to visit their primary care physician. However, because the majority of such symptomatic diagnoses present at later stages, the impact of disease detection through symptomatic presentation alone on cancer mortality remains modest [3]. Early screening of at-risk persons would allow timely identification of lung cancer cases, ideally at the pre-symptomatic stage when the tumour may still be resectable [4]. The two established screening technologies are chest X-ray (CXR) and low-dose computed tomography (LDCT) [5].

In recent years several clinical trials comparing imaging options have been performed. In the Prostate, Lung, Colorectal, and Ovarian (PLCO) trial, annual CXR provided no additional benefit compared to regular care [6]. The 2011 National Lung Screening Trial (NLST) compared annual LDCT to annual CXR in a high risk population. It showed a 20% reduction in lung cancer related mortality after three annual LDCT screenings [7]. Two large clinical trials are expected to present results in 2016. The first of these is the Dutch-Belgian Randomized Lung Cancer Screening Trial (NELSON), comparing LDCT to no screening in a population of 15,822 high risk individuals [8]. The second is the UK Lung Screen (UKLS) study which randomised 28,000 high risk persons to either LDCT or usual care [9].

A number of studies have sought to establish the cost-effectiveness of lung cancer screening since Black *et al*'s 2006 systematic review found insufficient evidence on the long term survival of screened patients to conduct a rigorous analysis [10]. The same authors in 2015 found that the CT screening arm of the NLST trial was associated with a higher cost (roughly £990) and 0.0201 more quality-adjusted life years (QALYs) compared to no screening, yielding an incremental cost-effectiveness ratio (ICER) of roughly £49,000 per QALY [11]. This result was highly sensitivity to a number of factors, including the population analysed and the ability to reproduce the results seen in the trial. Pugina *et al*'s 2015 review highlights the development made in the literature since 2006, finding nine studies of relevance [12]. The authors found that seven of the studies judge screening to be cost-effective when considered against a willingness-to-pay threshold of £66,000 per QALY and five at a lower threshold of £33,000 per QALY, thresholds of cost-effectiveness often

used in the US and Canada. Finally, Field *et al.* used a risk prediction tool to select persons for screening and (assuming a 20% mortality reduction based on the NLST), estimated the cost-effectiveness of screening to be roughly £8500 per QALY [9]. While the studies can be argued to be highly disparate in nature, evaluating different screening methodologies from different perspectives, they undoubtedly show the potential for cost-effectiveness of a LDCT-based lung screening program.

Despite these promising outcomes, there is still no global adoption of national lung cancer screening programs. The only exception has been the US where lung cancer screening is reimbursed for high risk persons since 2015 [13].

In this paper we argue that a demonstration of effectiveness and cost-effectiveness of LDCT screening to identify lung cancer at an earlier stage is only one hurdle in the decision to adopt national screening programs. We identify the associated challenges of screening capacity and low uptake as additional hurdles that must be overcome. We conclude by proposing a two-tier screening approach, looking to existing cancer screening programs and new technologies for inspiration.

The capacity burden of a screening program

The existing literature identifies that due to the significant health gains associated with early diagnosis of lung cancer and the relatively small marginal cost of each additional LDCT screen, there is potential for a LDCT screening program to be cost-effective. However, this is tempered by the large number of false positives (96.4% in NLST, but expected to be lower for the NELSON and UKLS studies) with associated cost and health outcome implications. A further issue is the uncertainty in the reproducibility and transferability of trial findings.

A key barrier in policy uptake can be argued to be the high CT capacity demanded by a LDCT screening initiative. Using the NLST eligibility criteria (age 55-74, at least 30 pack-year smoking history, former smokers quit within the last 15 years) applied to a UK setting we estimate a population of 677,491 would be eligible for every cycle of a screening program in the UK, which may occur at annual intervals. This is comparable with the 627,759 single (imaging) area, no contrast CTs conducted in 2013-14 in the English NHS [14]. As such challenges of implementation neither fall within the clinical nor cost-effectiveness evaluations, it is a hurdle that has been largely overlooked.

Associated with this capacity issue are the challenges associated with defining a high risk population to be screened. These can be seen as two-fold. Firstly, the definitions used in many trials (and exemplified by the NLST eligibility criteria) are broad, and thus result in a low yield of lung cancer ‘hits’ as a percentage of total screens (3.58% in NLST [7] and 3.75% in an early cut of UKLS [15]). Secondly, the definitions are poor at capturing the majority of lung cancers, with research suggesting that the NLST criteria would only capture 26.7% of lung cancer cases in the US [16]. Any attempt to broaden the definition used to capture more cancers would further increase the demands on LDCT capacity. This impact may be diminished if risk prediction tools are used to select high risk persons for screening [17].

The case for a two-tiered screening policy

Arguably the key difference between screening policies in lung cancer and existing screening programmes in other cancer areas is the presence of a cheap but effective first-line screening technology. For example, the colorectal cancer screening programmes in the NHS and the Netherlands makes use of postal Faecal Occult Blood Testing (FOBT) kits. The advantage of having a first tier screening technology is two-fold. Firstly, they represent an inexpensive way of identifying a high risk sub-group for further investigation without relying on symptomatic presentation [3], weakly correlated risk factors, or needing a very large standing colonoscopy/CT capacity. Secondly, a first-line screening technology may also present the eligible population with a much lower perceived barrier to participate in the screening program. Such a barrier directly impacts the cost-effectiveness of a nation-wide screening program by effectively reducing the size of the screened population [18]. The impact of non-adherence on cost-effectiveness is exacerbated if it occurs mostly in particular subgroups, for example those with low health literacy.

Several studies have identified factors predictive of adherence. Factors preventing adherence in the UKLS trial were mostly practical (47%) or emotional (18%). Furthermore, adherence was significantly lower in high risk groups, for instance among current smokers (Odds Ratio, OR: 0.70, 95% confidence interval (CI) of 0.63 to 0.78), older people (OR 0.73, 95% CI of 0.64 to 0.80) and lower socioeconomic groups (OR 0.56, 95% CI of 0.49 to 0.65) [19]. Stated preference studies for cancer types other than lung cancer found that although many in the target populations preferred screening above no screening [20,21], a significant percentage would rather not be screened for a variety of reasons [22,23].

Both the retrospective studies and stated preference studies further the rationale for why a first-line screen may be highly effective in cancer screening: it reduces barriers to entry in the screening programme at a low cost. These advantages are critical for lung cancer screening because adherence barriers seem to be more relevant for the high risk groups and because the large initial costs of CT scanners may be prohibitive for a nationwide screening policy.

New lung cancer screening technologies currently under development

Breath analysis

There are two (families of) technologies that may play the role of a first-line lung cancer screening technology: breath analysis and blood biomarkers. Breath analysis measures volatile organic compounds in the exhaled air with a device similar to a lung function monitor [24]. Different patterns of such compounds can differentiate between lung cancer patients and controls [25]. Adiguzel and Kulah reviewed nine studies and estimated sensitivity between 71% and 100% , and specificity between 73% and 90% [25]. However, the total sample size of the cited studies is small and the controls were mostly lung patients that have a high lung cancer risk (e.g. had COPD or bronchitis). Furthermore, differentiating between cancer stages is still an unresolved question, as is the best manner of obtaining breath samples. Since the measurement procedure is similar to a lung function test, it is unlikely that there are side effects or safety issues.

Blood biomarkers

Biomarkers indicative of lung cancer may not always be able to pass from a person's blood into the exhaled air, and thus may not be detectable with breath analysis. However, collecting blood samples can yield access to such markers [26]. Biomarkers furthest along the research track are autoantibodies, serum-based micro RNA signatures, and plasma-based micro RNA signatures [27–29]. The sensitivity/specificity for these compared to LDCT are 41%/88%, 82%/90%, and 87%/81%, respectively with trial sample sizes of 2289, 147, and 939 [30]. Three other biomarkers currently being piloted in trials are Ciz-1, circulating tumour cell traps, and EarlyCDT-Lung [31–33]. Sozzi and Boeri list six unresolved barriers in

the area of biomarkers: unreliable/inefficient discovery methods due to rapidly changing technology, genetic heterogeneity in tumours, reproducibility of laboratory assays, a lack of high-quality clinical trials, low reproducibility of results due to low concentration of analytes, and the lack of prospective biological samples [30]. A further issue is the specificity of biomarkers with regard to tumour type, which would require further diagnostics to identify the location of the tumour and treatment strategy. These technologies are still early in development and independent comparative trials are currently lacking, but the initial research results appear promising.

Discussion and future research

The idea of a two-tier screening policy has been applied in other fields of cancer. FOBt is a low cost method that can be done by people from the comfort of their own homes. X-ray, although requiring a visit to the hospital and lacking comfort, is arguably also an inexpensive first-line screening technology for breast cancer screening. For lung cancer, Sozzi recommended a similar approach by advising blood biomarker analyses in addition to the LDCT screen [30]. They calculate that such an approach would reduce the false positive rate from 19% to 4%.

Although the sensitivity and specificity is lower in the proposed first-tier screening tools than that of LDCT, it is important to note that factors other than diagnostic accuracy may be relevant for a first-line screening technology. First of all, the costs should be low to enable economically sustainable nation-wide use. Although both breath analysis and blood biomarkers are still early in development, we expect that the costs to be several orders of magnitude lower than that of a CT scanner, due to lower unit cost and limited infrastructure requirements. Blood biomarkers have the cost advantage that it may be possible to include the lung cancer markers into regular blood checkups. Secondly, it is important that the barrier to the first-line screen is low. We suspect that the possibility of implementing breath analysis and blood biomarkers in a primary care setting will lower the barriers to entry for eligible persons. Thirdly, it may be argued that specificity is a more important characteristic for a first-line screening technology than sensitivity since LDCT is already highly sensitive. A highly specific first-line screen in practice would mean that persons with negative test results can go home, while those with positive or unclear results can be scheduled for an LDCT. Additionally, it is important to not look at lung cancer screening in isolation. Screening is but one strategy in a policy maker's toolbox of reducing lung cancer mortality. Other strategies are smoke cessation programmes, awareness

campaigns, treatment, and recurrence monitoring. Arguably the most cost-effective overall strategy is some combination of these.

The discussion of lung cancer screening in this paper has pointed to several areas for further research. First of all, for any technology to be considered as a first-line screen, sufficient clinical evidence needs to be gathered. There is a need for (independent) comparative trials in populations other than only very high risk persons. Secondly, more insight is required into the natural progression of lung cancer to be able to more accurately determine what populations should be screened and when. Thinking about the new screening technologies as part of a two-tier screening programme may imply research questions not directly relevant when thinking about the new technologies as potential LDCT-replacements. Instead of focussing on proving that a technology is superior to LDCT, research could focus on proving that a LDCT programme *complemented* with a first-line screening technology outperforms a programme with *just* LDCT. Since the two main advantages of a first-line screening technology are low costs and increased adherence, more research is required to identify (and lower) costs, and to identify factors driving screening adherence. If then, for example, it is found that the ability to screen at a primary care facility is important, developers could try and design a technology that is sufficiently straightforward to be used by a GP. It is also important to find out if there is preference heterogeneity in the at risk population, for example preference subgroups that value different aspects of screening. It would be valuable to know if there are identifiable person characteristics that may help identify persons in preference subgroups that are likely to be less policy-adherent [23]. Such knowledge may, for example, be used for more targeted policy information strategies to areas with specific demographic characteristics. Finally, the evaluation, comparison and design of holistic cancer strategies may require more complex simulation methods [34].

Conclusion

In this paper we have examined the cost-effectiveness of lung cancer screening using LDCT, and tried to answer the question of why, despite promising clinical and health economic characteristics, LDCT-based screening does not seem to have taken off as a nation-wide healthcare policy. We argue that failure to consider the large CT capacity burden a LDCT-based screening would require and the poor levels of adherence to CT screening (especially in high-risk groups) have played major roles. We propose a two-tier screening policy where the first-tier is formed by a low cost technology with low barriers to entry, and a second tier

formed by LDCT. We argue that such an approach may increase screening adherence and reduce the cost burden of a nationwide screening policy. We have discussed several potential candidates for a first-line technology as well as their (dis)advantages.

Acknowledgements

The authors would like to thank Dr. Matthew Callister for his valuable comments on earlier drafts of this paper.

References

- [1] J.R. Molina, P. Yang, S.D. Cassivi, S.E. Schild, A. a. Adjei, Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship, *Mayo Clin. Proc.* 83 (2008) 584–594.
- [2] Cancer Research UK statistics, (2015). <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type>.
- [3] A.E. Ades, M. Biswas, N.J. Welton, W. Hamilton, Symptom lead time distribution in lung cancer: natural history and prospects for early diagnosis., *Int. J. Epidemiol.* 43 (2014) 1865–1873.
- [4] D. Sharma, T.G. Newman, W.S. Aronow, State of the art paper Lung cancer screening : history , current perspectives , and future directions, *Arch. Med. Sci.* 11 (2015) 1033–1043.
- [5] H.-U. Kauczor, L. Bonomo, M. Gaga, K. Nackaerts, N. Peled, M. Prokop, M. Remy-Jardin, O. Von Stackelberg, J.-P. Sculier, H.-U. Kauczor, ESR/ERS white paper on lung cancer screening, *Eur Respir J.* 46 (2015) 28–39.
- [6] M.M. Oken, W.G. Hocking, P. a. Kvale, G.L. Andriole, S.S. Buys, T.R. Church, E.D. Crawford, M.N. Fouad, C. Isaacs, D.J. Reding, J.L. Weissfeld, L. a. Yokochi, B. O'Brien, L.R. Ragard, J.M. Rathmell, T.L. Riley, P. Wright, N. Caparaso, P. Hu, G. Izmirlian, P.F. Pinsky, P.C. Prorok, B.S. Kramer, A.B. Miller, J.K. Gohagan, C.D. Berg, Screening by Chest Radiograph and Lung Cancer Mortality, *JAMA J. Am. Med. Assoc.* 306 (2011) 1865–1873.
- [7] The national lung screening trial research team, Reduced lung-cancer mortality with low-dose computed tomographic screening, *N. Engl. J. Med.* 365 (2011) 395–409.
- [8] M.A. Heuvelmans, M. Oudkerk, G.H. De Bock, H.J. De Koning, X. Xie, P.M. a Van Ooijen, M.J.W. Greuter, P. a. De Jong, H.J.M. Groen, R. Vliegthart, Optimisation of volume-doubling time cutoff for fast-growing lung nodules in CT lung cancer screening reduces false-positive referrals, *Eur. Radiol.* 23 (2013) 1836–1845.
- [9] J.K. Field, S.W. Duffy, D.R. Baldwin, D.K. Whynes, A. Devaraj, K.E. Brain, T. Eisen, J. Gosney, B. a Green, J. a Holemans, T. Kavanagh, K.M. Kerr, M. Ledson, K.J. Lifford, F.E. McDonald, A. Nair, R.D. Page, M.K.B. Parmar, D.M. Rassl, R.C. Rintoul, N.J. Screatton, N.J. Wald, D. Weller, P.R. Williamson, G. Yadegarfar, D.M. Hansell, UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening., *Thorax.* (2015) 1–10.
- [10] C. Black, A. Bagust, A. Boland, S. Walker, C. McLeod, R. De Verteuil, J. Ayres, L. Bain, S. Thomas, D. Godden, N. Waugh, The clinical effectiveness and cost-effectiveness of computed tomography screening for lung cancer: systematic reviews., *Health Technol. Assess. (Rockv).* 10 (2006).
- [11] W.C. Black, I.F. Gareen, S.S. Soneji, J.D. Sicks, E.B. Keeler, D.R. Aberle, A. Naeim, T.R. Church, G. a. Silvestri, J. Gorelick, C. Gatsonis, Cost-Effectiveness of CT Screening in the National Lung Screening Trial, *N. Engl. J. Med.* 371 (2014) 1793–1802.

- [12] A. Puggina, A. Broumas, W. Ricciardi, S. Boccia, Cost-effectiveness of screening for lung cancer with low-dose computed tomography: a systematic literature review., *Eur. J. Public Health.* 5 (2015) 1–8.
- [13] What screening tests are there?, *Cent. Dis. Control Prev.* (2015). http://www.cdc.gov/cancer/lung/basic_info/screening.htm.
- [14] NHS, NHS reference costs 2013 to 2014, *NHS Ref. Costs.* (2015). <https://www.gov.uk/government/publications/nhs-reference-costs-2013-to-2014>.
- [15] J.K. Field, D. Baldwin, S. Duffy, J. Holemans, M. Ledson, N. Screatton, R.C. Rintoul, D.M. Hansell, A. Devaraj, UK Lung Cancer Screening Trial: – First 1000 recruits, in: 8th NCRI Cancer Conf., Liverpool, 2012. <http://conference.ncri.org.uk/abstracts/2012/abstracts/A116.html>.
- [16] R.J. Optican, C. Chiles, Implementing lung cancer screening in the real world: opportunity, challenges and solutions, *Transl. Lung Cancer Res.* 4 (2015) 353–364. doi:10.3978/j.issn.2218-6751.2015.07.14.
- [17] S. a Kovalchik, M. Tammemagi, C.D. Berg, N.E. Caporaso, T.L. Riley, M. Korch, G. a Silvestri, A.K. Chaturvedi, H. a Katki, Targeting of low-dose CT screening according to the risk of lung-cancer death., *N. Engl. J. Med.* 369 (2013) 245–54. doi:10.1056/NEJMoa1301851.
- [18] P.M. McMahon, R. Meza, S.K. Plevritis, W.C. Black, C.M. Tammemagi, A. Erdogan, K. Ten Haaf, W. Hazelton, T.R. Holford, J. Jeon, L. Clarke, C.Y. Kong, S.E. Choi, V.N. Munshi, S.S. Han, J. Van Rosmalen, P.F. Pinsky, S. Moolgavkar, H.J. De Koning, E.J. Feuer, Comparing benefits from many possible computed tomography lung cancer screening programs: Extrapolating from the National Lung Screening Trial using comparative modeling, *PLoS One.* 9 (2014).
- [19] N. Ali, K.J. Lifford, B. Carter, F. McDonald, G. Yadegarfar, D.R. Baldwin, D. Weller, D.M. Hansell, S.W. Duffy, J.K. Field, K. Brain, Barriers to uptake among high-risk individuals declining participation in lung cancer screening: a mixed methods analysis of the UK Lung Cancer Screening (UKLS) trial, *BMJ Open.* 5 (2015) e008254.
- [20] S. Wortley, G. Wong, A. Kieu, K. Howard, Assessing stated preferences for colorectal cancer screening: a critical systematic review of discrete choice experiments, *Patient.* 7 (2014) 271–282.
- [21] E.W. de Bekker-Grob, J.M. Rose, B. Donkers, M.-L. Essink-Bot, C.H. Bangma, E.W. Steyerberg, Men's preferences for prostate cancer screening: a discrete choice experiment., *Br. J. Cancer.* 108 (2013) 533–41.
- [22] K.A. Phillips, S. Van Bebbber, D. Marshall, J. Walsh, L. Thabane, A review of studies examining stated preferences for cancer screening., *Prev. Chronic Dis.* 3 (2006) A75.
- [23] J.M. Hummel, L.G.M. Steuten, C.J.M. Groothuis-Oudshoorn, N. Mulder, M.J. Ijzerman, Preferences for colorectal cancer screening techniques and intention to attend: A multi-criteria decision analysis, *Appl. Health Econ. Health Policy.* 11 (2013) 499–507.
- [24] H.P. Chan, C. Lewis, P.S. Thomas, Exhaled breath analysis: novel approach for early detection of lung cancer., *Lung Cancer.* 63 (2009) 164–8.
- [25] Y. Adiguzel, H. Kulah, Breath sensors for lung cancer diagnosis, *Biosens. Bioelectron.* 65 (2015) 121–138.
- [26] D. Xiang, B. Zhang, D. Doll, K. Shen, G. Kloecker, C. Freter, Lung cancer screening: from imaging to biomarker., *Biomark. Res.* 1 (2013) 4.
- [27] C.J. Chapman, A. Murray, J.E. McElveen, U. Sahin, U. Luxemburger, O. Türeci, R. Wiewrodt, A.C. Barnes, J.F. Robertson, Autoantibodies in lung cancer: possibilities for early detection and subsequent cure., *Thorax.* 63 (2008) 228–233.
- [28] F. Bianchi, F. Nicassio, M. Marzi, E. Belloni, V. Dall'Olio, L. Bernard, G. Pelosi, P. Maisonneuve, G. Veronesi, P.P. Di Fiore, A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer, *EMBO Mol. Med.* 3 (2011) 495–503.

- [29] M. Boeri, C. Verri, D. Conte, L. Roz, P. Modena, F. Facchinetti, E. Calabro, C.M. Croce, U. Pastorino, G. Sozzi, MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer, *Proc. Natl. Acad. Sci.* 108 (2011) 3713–3718.
- [30] G. Sozzi, M. Boeri, Potential biomarkers for lung cancer screening., *Transl. Lung Cancer Res.* 3 (2014) 139–48.
- [31] G. Higgins, K.M. Roper, I.J. Watson, F.H. Blackhall, W.N. Rom, H.I. Pass, J.F.X. Ainscough, D. Coverley, Variant Ciz1 is a circulating biomarker for early-stage lung cancer., *Proc. Natl. Acad. Sci. U. S. A.* 109 (2012) E3128–35.
- [32] Z. Zhang, N. Ramnath, S. Nagrath, Current Status of CTCs as Liquid Biopsy in Lung Cancer and Future Directions, *Front. Oncol.* 5 (2015) 1–11.
- [33] S. Lam, P. Boyle, G.F. Healey, P. Maddison, L. Peek, A. Murray, C.J. Chapman, J. Allen, W.C. Wood, H.F. Sewell, J.F.R. Robertson, EarlyCDT-Lung: An immunobiomarker test as an aid to early detection of lung cancer, *Cancer Prev. Res.* 4 (2011) 1126–1134.
- [34] D. A. Marshall, L. Burgos-liz, M.J. IJerman, N.D. Osgood, W. V Padula, M.K. Higashi, P.K. Wong, K.S. Pasupathy, W. Crown, Applying Dynamic Simulation Modeling Methods in Health Care Delivery Research — The SIMULATE Checklist: Report of the ISPOR Simulation Modeling Emerging Good Practices Task Force, *Value Heal.* 18 (2015) 5–16.

Chapter 6

Public preferences for lung cancer screening policies

Henk Broekhuizen

Catharina G.M. Groothuis-Oudshoorn

Rozemarijn Vliegenthart

Harry J.M. Groen

Maarten J. IJzerman

Abstract

Background and aims: Since early detection of lung cancer can substantially influence overall survival, there is increasing attention for lung cancer screening. This study aims to identify public preferences for lung cancer screening and to identify preference subgroups.

Methods: The study was designed as a multi-attribute elicitation experiment using swing weighting. Attributes were selected using interviews and a panel session. Hierarchical clustering was used to identify preference subgroups.

Results: 1034 respondents from a representative Dutch panel aged between 40 and 80 were included. The identified attributes were location of screening (weight=0.18, SD=0.16), mode of screening (0.17, SD=0.14), sensitivity (weight=0.16, SD=0.13), specificity (weight=0.13, SD=0.12), waiting time until results (weight=0.13, SD=0.12), radiation load (weight=0.13, SD=0.12), and duration of screening procedure (weight=0.10, SD=0.09). Most respondents preferred breath analysis (45%) to blood samples (31%) or a scanner (24%). 59% would prefer to be screened at their GP instead of the hospital. There was a significant difference in education between clusters ($P<0.01$). There were no significant differences between clusters regarding gender, age, smoke status, self-perceived risk, or actual cancer risk.

Conclusion: There is much variation in what value people attach to attributes of lung cancer screening. Our results indicate that there is great potential for new screening technologies that can be used at a primary care facility, and that a one-size-fits-all approach for lung cancer screening is unlikely to provide the best value for the screening population.

Introduction

Lung cancer is one of the leading causes of cancer-related deaths in the world [1]. Despite recent advances in treatment options, overall survival remains low [2]. Since early detection of (non-small cell) lung cancer can substantially decrease overall cancer mortality, there is increasing attention for developing lung cancer screening policies [3]. A first choice in designing such a policy is to choose a screening technology. Most research into screening for lung cancer has been done for imaging technologies, where it was found that chest x-ray provided no additional survival benefit over sputum cytology [4] and that low-dose computed tomography (LDCT) reduced cancer-related mortality by 20% compared to chest x-ray [5,6]. Less expensive screening technologies with a lower barrier to entry may be found in breath or blood biomarkers [7,8]. However, since these technologies are early in development there are still issues surrounding accuracy, standardization, calibration and method of sample collection.

Understanding the preferences of the eligible population for characteristics of screening policies may help policy makers choose optimal attributes of screening policy that maximize screening attendance and it may help to tailor information provision for people who are eligible for screening [9]. Achieving high attendance (particularly of high risk subpopulations) is important to realize the reduction in cancer-related mortality promised by cancer screening trials. Studies performed alongside lung cancer screening trials reported that smoking status, perceived risk of lung cancer and result of baseline CT scan were indicative of adherence to the screening protocol [10–14]. To obtain quantitative insights into the relative importance of attributes of lung cancer screening, stated preference methods can be used [9,15,16]. Stated preference questionnaires present respondents with direct questions about their preferences or have respondents make hypothetical decisions from which preferences are subsequently inferred. These types of methods are increasingly used in healthcare [15,17] but no results from stated preference studies for lung cancer screening are known [18]. There have, however, been stated preference studies in the context of screening for other types of cancer: namely breast [16], colorectal [9,19–22] and prostate [23–25] cancer. Common findings in these studies were that the most important attribute according to respondents was test accuracy, and that respondents indicated a higher willingness to attend if they had a higher self-perceived risk of cancer. It was also found that there can be distinct differences in weights assigned to screening attributes by (groups of) people [25]. Knowledge on this so-called preference heterogeneity may be useful because it would allow for concrete targets for motivating specific subgroups to attend a screening. Furthermore, making decisions while taking into

account preference subgroups may be a step toward more personalized healthcare [26,27], and may improve shared decision making between health provider and screening program attendees.

To date there have been no studies that investigate the preferences of people for attributes of lung cancer screening. Furthermore, there is no knowledge of preference subgroups in that part of the population that is eligible for lung cancer. The aims of this paper were therefore to measure the preference weights for attributes of lung cancer screening policies in a screening-eligible population and to empirically identify preference subgroups.

Methods

Study population

A public sample from the general population in the Netherlands was selected for this study through an online panel bureau (SSI). Men and women aged between 40 and 80 years who had no (history of) lung cancer were included.

Interviews and panel session to establish attributes and levels

To identify attributes, exploratory interviews were conducted with three clinical experts (one MD/PhD student, one lung oncologist and one radiologist) and with two persons from the intended study population. A one-day panel session was organized in which another ten respondents from the intended study population identified attributes they considered relevant. This was done by each participant writing down attributes he/she considered important on a post-it note, after which a group discussion took place where the group as a whole worked on remove overlapping attributes. The authors then presented the findings from the interviews, and together with the panelists a final list of attributes was defined by sorting the attributes from most important to least important and picking the seven most important attributes. The selected attributes for inclusion in the questionnaire were sensitivity, specificity, radiation load, duration of screening procedure, time until results of the screening are communicated, mode of screening, and location of screening. *Sensitivity* is the probability of a positive (i.e. suspect) test result given that someone has lung cancer. *Specificity* is the probability of a negative (i.e. non-suspect) test result given that someone does not have lung cancer. *Radiation load* was defined as the radiation that someone receives during a single screening on top of the normal yearly background radiation in the Netherlands [28].

	Attribute name	Worst level	Best level
Continuous attributes	Sensitivity	70%	100%
	Specificity	70%	100%
	Radiation load	Background + 1.5 mSv	Background
	Duration of screening procedure	45 minutes	15 minutes
	Time until results	14 days	1 day
Categorical attributes	Scan type	levels: 'lie in scanner', 'sustained breath into device', 'give blood'	
	Location	levels: 'nearest hospital', 'at your GP'	

Table 1: Included attributes and levels. Sensitivity is the probability of a positive/suspect test results given that the person has lung cancer. Specificity is the probability of a negative test results given that the person does not have lung cancer. The worst and best levels for the continuous attributes were defined based on interviews with clinicians. For the categorical attributes the preference order was elicited from respondents in separate questions.

The *duration of screening* was defined as the time the screening attendee would spend at the facility where screening takes place and was measured in minutes. *Time until results* was defined as the time it takes until the screening attendee receives the results of the screening either through a consultation or with a letter sent to his/her home address. It was measured in days. *Mode of screening* was defined as the personal experience of someone being screened, i.e. what they have to undergo in order to be screened. The levels for this attribute were 'lie in scanner', 'breath into device', and 'give blood', corresponding to LDCT, breath analysis and blood-based screening [29–31]. Finally, *location of screening* is the where screening takes place and could be either 'at closest hospital' or 'at the office of your general practitioner'. For interpretation purposes, each attribute was assigned a lower ('worst') and an upper ('best') level. The included attributes and their levels are presented in Table 1.

Measuring attribute weights with swing weighting

For the measurement of preferences a multi-attribute utility theory (MAUT) framework was adopted [32]. This theory states that people choose between services based on the value of such services, and that the value of services can be decomposed into a set of attributes. The relative importance of each attribute is denoted by an attribute weight. The interpretation of an attribute weight in MAUT is the relative importance of improvements from the 'worst' level to the 'best' level between attributes [33]. This necessitates that these are defined for each of the attributes. The worst and best performance levels for the

attributes sensitivity, specificity, and radiation load were determined based on clinical literature [29–31] and was confirmed with the clinicians in interviews. The worst and best attribute levels for the other attributes were determined in the panel session. This was done by each panelist writing down what he/she considered to be the worst and best level and then taking the lower and upper bounds of those. Since the preferential order for the levels of the location and mode attributes could not be assumed *a priori* or be decided in the panel session, we included separate questions in the questionnaire asking respondents to rank the mode and location levels from best to worst.

The swing weighting method was used in this study to elicit attribute weights from respondents [34,35]. In short, in our swing weighting procedure respondents were presented with a hypothetical screening policy that has the worst performance on all attributes. They were then asked to 1) rank the order in which they would like to improve attributes (i.e. change the performance from the worst to the best level) and 2) to rate the relative value of these improvements. For the latter, respondents were tasked to assign a number between 0 and 100 to each of the improvements. The rating for the improvement that was ranked first was fixed to 100. So, if a respondent feels a particular improvement is half as important as the improvement they ranked first, they will assign a rating of 50 to that particular improvement.

Questionnaire design and testing

The questionnaire consisted of three parts. Apart from general demographics, respondents were asked to provide information regarding the risk factors of the Liverpool Lung Project model [36]. Respondents were also asked if they considered their risk of lung cancer below average, average, or above average. The descriptive questions were followed by the swing weighting exercise described in the previous paragraph. The questionnaire was piloted in six face-to-face interviews. Since some respondents in the pilot tests had difficulty grasping the concepts of sensitivity and specificity, it was decided that these concepts would be explained in the questionnaire using four short descriptions with concrete numbers. The questionnaire was constructed in the Dutch language using LimeSurvey and hosted on the University of Twente Datalab servers. An English translation of the questionnaire can be found as supplementary material for the online version of this study.

Statistical analyses

Descriptives and lung cancer risks

The characteristics of the respondents in the sample were summarized using descriptive statistics. We divided the educational level of respondents in three categories: low denoted

primary or secondary education, medium denoted intermediate vocational training, and high denoted higher vocational training or university. The 5-year lung cancer risk for each respondent was calculated with the Liverpool Lung Project model [36].

Analyzing the attribute weights

We calculated the attribute weights from the rating given by respondents as follows. First of all, we excluded respondents who did not seem to engage with the swing weighting questions, i.e. those who did not give the highest rating to the attribute they ranked first (inconsistent answer) and those who gave a rating of zero for all attributes except their first ranked attribute (indicative of skipping the question as the rating for the first ranked attribute was fixed). Then, the attribute weight for each attribute k according to a particular respondent q was found by dividing the rating assigned to that attribute by the sum of ratings assigned by respondent q .

The resulting attribute weight is thus a normalized rating. It represents the value of an improvement from the worst level to the best level on that attribute, relative to such improvements for other attributes [37]. One sample T tests were used to test if the mean attribute weights (of the whole sample) differed significantly from $1/7 \approx 0.14$. This threshold was chosen because $1/7$ would be the attribute weight for each attribute if all attributes were equally important. A significant deviation from $1/7$ thus suggest that an attribute matters to people.

Cluster analysis methodology

To identify preference subgroups, a cluster analysis method was used. Cluster analysis is part of the family of unsupervised statistical learning methods [38]. Such methods are used to identify latent patterns in the data, as opposed to *a priori* cluster definition based on observable respondent characteristics. Because the attribute weights for the seven attributes add up to one per respondent, there were six degrees of freedom and we performed the cluster analysis on six attributes excluding the attribute with the lowest mean weight across the sample. Different clustering methods exist. In our study a hierarchical clustering method based on the correlation distance was used because the attribute weights per respondent are correlated (because they are normalized ratings) [39]. A second important choice in cluster analysis is the number of clusters. This should be large enough to be of sufficient clustering quality, but small enough to be interpretable [25]. A clustering solution is of good quality when the variation between respondents in the same cluster is minimized and when the variation between respondents in different clusters is maximized. This quality was measured in this study with the Dunn index [40]. For

interpretability reasons, we set the minimum size of a cluster to be 10% of the sample and decided that we would include no more than ten clusters. Since the interpretation of the preference clusters may differ depending on what levels respondents selected as ‘worst’ and ‘best’ levels for mode and location, we compared what levels of these were ranked first by respondents in each cluster.

We tested if the mean attribute weights across the clusters were significantly different using ANOVA tests. Furthermore, we used χ^2 tests to investigate if the following characteristics were significantly different between clusters: gender, percentage of smokers, percentage of respondents who had an above-average self-perceived risk of cancer, percentage of respondents who prefer a particular mode of screening, and percentage of respondents who prefer a particular location of screening. An ANOVA test was used to test if age and mean 5-year lung cancer risks differed significantly between clusters.

Sample size calculations

To be able to detect an effect size of 0.15 with a power of 80% at a significance level of 5% for the ANOVA test for differences between attribute weights over preference subgroups, a sample size of 70 is needed per subgroups. Since the number of subgroups (i.e. preference clusters) for the ANOVA was not known prior to data collection, a conservative estimate of ten groups was used. This resulted in a minimum required sample size of 704. A post-hoc Bonferroni p-value correction for multiple hypothesis testing was used.

Results

Respondent characteristics

In total, 1111 of the 1586 invited respondents completed the questionnaire (70%). Of these, ten were outside the age range of 40 to 80 years and were excluded from the sample. Fifty-one percent of the respondents in the sample were female, and the mean age was 58 years (SD = 11).

The respondents were split about evenly over the low, medium and high education categories. Twenty-three percent of the sample was a current smoker, while 37% were previous smokers, and 40% were non-smokers. More than half of the respondents in the sample (54%) had experience with lung cancer in their close social circle. The mean 5-year lung cancer risk in the sample was 1.8%. More detailed information about the sample characteristics is presented in Table 2.

<i>Characteristic</i>	<i>Mean (SD) or n (percentage)</i>
<i>Age (years)</i>	58 (11)
<i>Gender (female)</i>	526 (51%)
<i>Education</i>	
... <i>Low*</i>	323 (31%)
... <i>Medium†</i>	380 (37%)
... <i>High‡</i>	331 (32%)
<i>Self-perceived risk of lung cancer</i>	
... <i>Below average</i>	360 (35%)
... <i>Average</i>	561 (54%)
... <i>Above average</i>	113 (11%)
<i>Smoke status</i>	
... <i>Current smoker</i>	234 (23%), with mean smoking years 33 (SD=13)
... <i>Previous smoker</i>	387 (37%), with mean smoking years 22 (SD=13) and mean years since quitting 20 (SD=14)
... <i>Never smoker</i>	413 (40%)
<i>Current respiratory problems</i>	194 (18%)
<i>Has had pneumonia</i>	275 (26%)
<i>Previous tumor (not lung)</i>	71 (7%)
<i>Asbestos exposure</i>	49 (5%)
<i>Close social circle LC experience</i>	549 (54%)
... <i>Family member < 60 years old</i>	92 (9%)
... <i>Family member > 60 years old</i>	167 (17%)
... <i>But no family relation</i>	285 (28%)
<i>Mean 5-year risk of lung cancer (with 95% CI)</i>	
... <i>Whole population</i>	1.8% (0.0% to 11.6%)
... <i>Current smokers</i>	2.7% (0.0% to 16.0%)
... <i>Previous smokers</i>	1.3% (0.0% to 16.0%)
... <i>Never smokers</i>	0.2% (0.0% to 1.5%)

Table 2: Characteristics of the 1034 respondents included in the analyses. *=primary or secondary education. †=intermediate vocational training. ‡ =higher vocational training or university education. CI= confidence interval.

<i>Attribute</i>	Rank order (1st > 2nd > 3rd)	n (%)	Corresponding attribute weight (SD)	p-value
<i>Mode</i>	Blood > Breath > Scanner	231 (22%)	0.19 (0.16)	<0.001
	Blood > Scanner > Breath	88 (9%)	0.18 (0.13)	
	Breath > Blood > Scanner	353 (34%)	0.14 (0.11)	
	Breath > Scanner > Blood	111 (11%)	0.16 (0.14)	
	Scanner > Blood > Breath	155 (15%)	0.20 (0.17)	
	Scanner > Breath > Blood	96 (9%)	0.17 (0.17)	
<i>Location</i>	GP > Hospital	611 (59%)	0.17 (0.14)	0.69
	Hospital > GP	423 (31%)	0.16 (0.15)	

Table 3: Sample-level rankings of the levels of categorical attributes. The third column describes how often each rank order occurred. The fourth column presents the attribute weights for respondents of each particular rank order (i.e. mode of screening for the first six rows, and location of screening for the bottom two rows). Finally, the fifth column shows the p-value for the ANOVA that tests if the attribute weights of either mode of location differed significantly between the rank orders.

Ranking of categorical attributes levels and weights of attributes

After collecting the answers to the SWING weighting questions from respondents, we identified 67 respondents who did not give the highest rating to the attribute they ranked first or who gave a rating of zero for all attributes except their first ranked attribute. These respondents were excluded before the final analyses. The characteristics of the excluded respondents did not differ significantly from those of the included respondents. The number of respondents in the final analyses (1034) was more than the required minimum number of respondents (704). Of all respondents included in the final analyses, 464 respondents (45%) preferred breathing into a device, 319 respondents (31%) preferred giving blood, and 251 respondents (24%) preferred laying in a CT-scanner. The most common preference ordering for mode of screening was breath > blood > scan and the least occurring rank order was blood > scanner > breath (Table 3). There were 609 respondents (59%) who would prefer to be screened at their GP office instead of at the hospital (Table 3). The attributes most often ranked first were location (27%), mode (20%), and sensitivity (19%). This is reflected in the attribute weights (i.e. the normalized ratings) of these attributes when averaged over the sample: location had a weight of 0.18 (SD = 0.16), mode had a weight of 0.17 (SD = 0.14), and sensitivity had a weight of 0.16 (SD = 0.13). These three most important attributes were followed by specificity (weight=0.13, SD=0.12), waiting time until results (weight=0.13, SD=0.12), and radiation load (weight=0.13, SD=0.12). Finally, duration of the screening procedure was the least

important attribute, ranked first by only 4% of respondents and with a mean weight of 0.10 (SD=0.09). The mean attribute weights for all but one attribute (radiation load) were significantly different from $\frac{1}{7}$. Table 3 shows that the attribute weight for mode of screening was significantly different between the various preference orders for mode of screening ($P<0.001$). The attribute weight for location of screening did not differ significantly between respondents who preferred the hospital location and respondents who preferred the GP location. An overview of the mean attribute weights is presented in Table 4 and Figure 1.

Preference subgroups

Based on the discussed requirements for cluster analysis, we included five clusters (Table 5). We assigned a name to each cluster based on the two attributes with the highest mean attribute weights according to respondents in that cluster. As a further condition we did not include an attribute in a cluster's name if its mean attribute weight did not differ significantly from $\frac{1}{7}$ in the cluster. The latter implies that a cluster could be assigned the name of only one attribute. The names of the five clusters as assigned in this manner were: 'radiation-sensitivity' (n=236), 'waiting time until results' (n=157), 'location-mode' (n=299), 'mode-sensitivity' (n=184), and 'sensitivity-specificity' (n=158). So for example, the 236 respondents in the 'radiation-sensitivity' cluster considered radiation load and sensitivity to be the most important attributes of screening. The mean attribute weights differed significantly between clusters for radiation load ($p<0.01$), duration of screening procedure ($p<0.01$), mode of screening ($p<0.01$), and location of screening ($p=0.01$).

Attribute name	Mean normalized weight (SE)
<i>Sensitivity</i>	0.16 (0.004) †
<i>Specificity</i>	0.13 (0.004)*
<i>Radiation load</i>	0.13 (0.004)
<i>Duration of screening procedure</i>	0.10 (0.003) †
<i>Waiting time until results</i>	0.13 (0.004) †
<i>Mode of screen</i>	0.17 (0.004) †
<i>Location of screen</i>	0.18 (0.005) †

Table 4: Sample-level normalized attribute weights. P values for the one sample t test to test if the mean weight is different from $\frac{1}{7}$ are denoted as *= $p<0.05$ and †= $p<0.01$.

		Radiation- sensitivity	Mode- sensitivity	Location- mode	Wait times	Sensitivity- specificity	p-value
	<i>n</i>	236	184	299	157	158	
Normalized criterion weights	<i>Sensitivity</i>	0.21 (0.12)	0.15 (0.11)	0.08 (0.06)	0.09 (0.06)	0.31 (0.21)	p=0.14
	<i>Specificity</i>	0.20 (0.11)	0.10 (0.06)	0.06 (0.04)	0.07 (0.05)	0.26 (0.17)	p=0.93
	<i>Radiation load</i>	0.25 (0.18)	0.10 (0.06)	0.10 (0.06)	0.14 (0.09)	0.07 (0.05)	p<0.01
	<i>Duration of screening procedure</i>	0.09 (0.08)	0.11 (0.11)	0.10 (0.08)	0.13 (0.12)	0.08 (0.07)	p<0.01
	<i>Waiting time until results</i>	0.07 (0.05)	0.11 (0.06)	0.11 (0.08)	0.30 (0.18)	0.10 (0.07)	p=0.93
	<i>Mode of screen</i>	0.11 (0.10)	0.32 (0.20)	0.18 (0.12)	0.14 (0.08)	0.09 (0.05)	p<0.01
	<i>Location of screen</i>	0.07 (0.05)	0.12 (0.07)	0.37 (0.18)	0.13 (0.07)	0.09 (0.06)	p=0.01
categorical criteria	<i>Screening mode</i>						p<0.01
	<i>Scanner</i>	17.4%	33.2%	23.7%	28%	21.5%	
	<i>Breath analysis</i>	58.1%	33.2%	38.5%	46.5%	49.4%	
	<i>Blood analysis</i>	24.6%	33.7%	37.8%	25.5%	29.1%	
	<i>Screening location</i>						p=0.91
	<i>GP</i>	56.4%	59.8%	64.5%	55.4%	54.4%	
	<i>Hospital</i>	43.6%	40.2%	35.5%	44.6%	45.6%	

Table 5: Normalized attribute weights and most preferred levels of categorical attributes per cluster. The reported p-value is for the ANOVA or χ^2 tests for testing if the mean weight or rank order of categorical attribute levels differs across clusters.

The rank order of levels for the mode of screening differed significantly between clusters ($p<0.01$). Fifty-eight percent of the 236 respondents in the radiation-sensitivity cluster preferred the breath analysis mode compared to 45% in the whole sample and this difference was significant ($p<0.05$). In contrast, 33% of the 184 respondents in the mode-sensitivity mode preferred the scanner mode. The pattern of location preferences did not differ significantly between the clusters ($p=0.91$). There were also no significant between-cluster differences with regard to gender, age, smoke status, self-perceived risk, or 5-year lung cancer risk. Level of education differed significantly between clusters ($p<0.01$). Of

respondents with a low or medium level of education, most respondents were part of the location-mode cluster. Of respondents with a high level of education, most respondents were part of the sensitivity-specificity cluster. An overview of respondent characteristics across the various clusters can be found in Table 6.

	Cluster					p-value
	Radiation-sensitivity	Mode-sensitivity	Location-mode	waitUntilR esults	Sensitivity-specificity	
<i>n</i>	236	184	299	157	158	
<i>Percentage female</i>	54.7%	55.4%	49.2%	44.6%	49.4%	p=0.93
<i>Age (SD)</i>	55.7 (10.6)	59.1 (11.2)	59.6 (11.7)	58.9 (11.5)	58.3 (10.8)	p=0.13
<i>Percentage smokers</i>	19.1%	23.4%	21.7%	25.5%	25.9%	p=0.93
<i>Percentage AASPR</i>	11.9%	13.0%	6.7%	8.9%	17.1%	p=0.10
<i>5-year lung cancer risk</i>	1.3%	1.9%	1.9%	2.0%	1.6%	p=0.93
<i>Education</i>						p<0.01
... <i>Low</i>	25.0%	29.9%	36.5%	38.2%	25.3%	
... <i>Medium</i>	33.1%	44.6%	34.8%	37.6%	36.1%	
... <i>High</i>	41.9%	25.5%	28.8%	24.2%	38.6%	

Table 6: Comparison of observable characteristics between preference clusters. The reported p values are for the t or χ^2 tests that test if there are (significant) differences between the characteristics of the preference clusters. AASPR=above average self-perceived risk.

Discussion

This study examined public preferences for attributes of lung cancer screening policies. Based on interviews and a panel session the included attributes were sensitivity, specificity, radiation load, duration of screening procedure, waiting time until results, mode of screening, and location of screening. Of all respondents, 45% preferred breath analysis over going through a scanner or having a blood sample taken, and 60% would prefer to be screened at their GP's office rather than at a hospital. The most important attributes according to the respondents were location of screening, mode of screening, and sensitivity.

Our results indicate that a one-size-fits-all approach for lung cancer screening is unlikely to provide the best value for all eligible persons in the population. Of the total respondent population, 59% would prefer a GP-based policy, and for 30% of respondents location was the most important attribute. This suggests that a hospital-based policy may provide a lower value for a significant percentage of people in the population. An interesting finding is that the lowest weight was given to duration of screening procedure. During the panel session the respondents clearly stated that they made a distinction between ‘useful time duration’ (going through a scanner, speaking to a clinician/nurse) and ‘useless time duration’ (predominantly waiting time) and indicated much less willingness to accept the latter. The low weight given by the respondents to duration of screening procedure may thus primarily reflect that people are willing to invest time in the screening, but care should still be taken that the waiting times are not too long.

The results from the cluster analysis can provide relevant insights for implementing screening policies. A first observation is that there seems to be a split between respondents that focus on aspects of healthcare delivery (‘location-mode’ and ‘waiting time until results’ clusters) and respondents that focus on clinical aspects of screening (‘radiation-sensitivity’ and ‘sensitivity-specificity’ clusters). That 44% of respondents (Table 5) seem mostly concerned with aspects of healthcare delivery is an importance piece of information since most studies looking into how to implement lung cancer screening have until now focused on clinical attributes of screening. A second observation is that our results can help understand what drives people’s understanding of the value of screening, which can help guide tailored information provision during implementation. For example, for respondents in the location-mode cluster a screening based on breath or blood analysis at the GP is most valuable for practical reasons of travel and burden of screening procedure. A similar screening policy may be valuable for respondents in the radiation-sensitivity cluster; although the reason behind this increased value may be different (being concerned over scanner radiation load). If a scanner-based policy is nevertheless adopted, different implementation strategies may be appropriate for these two groups. The mode-location group may require more clarity about what it is like to be in a scanner and an incentive that reduces their perceived burden of going to the hospital (for example travel cost reimbursement or screening during the weekends). To convince the radiation-sensitivity group to come to the same screening, more information about what (little) impact the radiation load in LDCT has for them may be needed. A final observation when interpreting the cluster analysis results is that in this study the only predictive factor for cluster membership was educational level. One might therefore imagine tailored information

provision according to the preference clusters that persons with a particular level of education most often belong to. For example, brochures aimed at persons with a higher level of education could emphasize information about clinical benefits and the magnitude of harms due to radiation load.

This study is (to the best of our knowledge) the first to measure public preference for lung cancer screening. Its main strengths are a large public sample and a broad set of attributes. This enabled insight into the variation in preferences that may exist in the study population (as indicated by the preference clusters). The broad set of attributes is important because the choice to attend a screening program can be argued to depend on person-specific attributes (e.g. self-perceived risk), technology-specific attributes (e.g. sensitivity), and healthcare delivery-specific attributes (e.g. location). Most preference studies in the field of cancer screening have focused on technology-specific attributes of screening but this may lead to unrealistic ideas about what drives screening attendance [18]. In our study we therefore measured the preferences of the general public for technology-specific attributes of screening as well as for healthcare delivery-specific attributes of screening. Furthermore, we investigated whether preferences for these attributes were related to demographics, self-perceived risk, and lung cancer risk factors. Lung cancer screening trials previously investigated person-specific attributes in relation with screening attendance and found that smoking status and self-perceived risk correlated with screening attendance [10–14], which would suggest that such factors are predictive of population preferences. However, we found no preference clusters with a significantly different proportion of smokers or persons with an above average self-perceived risk of lung cancer. The difference between our findings and those from the lung cancer screening trials may be explained by the fact that screening trials generally include only high-risk persons while our study considered a public sample. A common finding in preference studies for other types of cancer is that clinical attributes of screening are considered most important by respondents. This was not reflected in our sample as a whole, although it was the case for two clusters ('radiation-sensitivity' and 'sensitivity-specificity' clusters). A first explanation for this may be that not all previous preference studies took into account such attributes. Furthermore, since respondents in the clinically-oriented clusters had a higher level of education than those in other clusters, the difference may be explained by the fact that our public sample was split about evenly between the three educational categories. Finally, an explanation may be that respondents perceive overlap between attributes, for example perceiving going through a scanner as also having intrinsic value with regard to diagnostic accuracy. In the questionnaire the respondents were explicitly asked to consider the mode and location

attributes independently from preconceptions of sensitivity and specificity, but we cannot rule out this was misunderstood by some respondents.

Our study also had some limitations. First of all, it should be taken into account that stated preferences as elicited with this study may differ from the revealed preference for screening policy of respondents. Concrete examples of this can be found for colon cancer screening, where a review found that 66% to 88% stated they would be willing to attend, while actual attendance in the population was much lower (40% to 50%) [21,22]. Also, there is uncertainty in our results because of the inherent difficulty respondents may experience when translating their preferences into rankings or numbers. This may have been the case with the 67 respondents who we excluded based on inconsistent answers. Finally, because we did not include 'intention to attend' questions, we can only compare the value of screening policies across respondents, but we cannot predict absolute screening attendance.

Future research should be directed toward using the results presented in this study to design and compare lung cancer screening policies that optimize value for the eligible population. A critical part of such an undertaking would be to assess the costs of implementing and running lung cancer screening policies [41]. Since a CT-based national policy may be costly, new less-costly technologies such as breath analysis or biomarkers may be advantageous [7,8]. Our study suggests that the perceived value of such screening technologies can be substantial if these could be realized in a primary care setting. Breath and other biomarker technologies are still early in development, but if clinical equivalence to LDCT can be proven, it may be possible to operationalize a national screening program using only breath/biomarker analysis. If the new technologies cannot prove equal or superior diagnostic performance, tradeoffs have to be made between attendance and screening sensitivity. The reasoning here is that because of the voluntary nature of screening, even a highly sensitive screening technology does not add value if people do not attend. One might envision a two-line screening policy where the first line is formed by breath/biomarker analysis at a primary care facility which will not be perceived as a barrier by most people. People with clearly negative results on the breath/biomarker analysis could then be sent home, while people with suspect or indeterminate results could be sent to a second line screening with LDCT at a hospital. A similar two-line strategy is currently used in the Netherlands for colon cancer, where the first line is formed by fecal occult blood tests individuals can perform at home, and the second line is formed by colonoscopy; the latter being a sensitive procedure but an expensive and burdensome one that may deter people from attending if it were the only screening option. For such a two-line policy

to be a practical solution, further research is needed to ascertain the accuracy of the technology used in the first line. In the panel session it was mentioned that it would be beneficial to combine several types of screening. Designing a multi-indication screening is arguably even more difficult than designing a single-cancer screening, but may have great potential in terms of mortality reduction [42].

Conclusion

In conclusion, in our study we assessed the preferences of the general population regarding performance attributes of lung cancer screening policies. Mode of screening, location of screening, and sensitivity were found to be the most important attributes. Respondents preferred GP-based screening and breath analysis. Five distinct preference subgroups were identified. Of these, two subgroups (44% of respondents) were oriented more toward aspects healthcare delivery and two subgroups (38% of respondents) were oriented more toward clinical benefits and harms of screening. There was a significant difference in educational level between clusters. A one-size-fits-all approach for lung cancer screening is unlikely to provide the best value for everyone in the eligible population.

References

- [1] Molina JR, Yang P, Cassivi SD, Schild SE, Adjei A a. Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clin Proc* 2008;83:584–94.
- [2] Siegel R, Ma J, Zou Z, Jemal A. Cancer Statistics , 2014. *CA Cancer J Clin* 2014;64:9–29.
- [3] Moyer VA. Screening for lung cancer: U.S. preventive services task force recommendation statement. *Ann Intern Med* 2014;160:330–8.
- [4] Oken MM, Hocking WG, Kvale P a., Andriole GL, Buys SS, Church TR, et al. Screening by Chest Radiograph and Lung Cancer Mortality. *JAMA J Am Med Assoc* 2011;306:1865–73.
- [5] The national lung screening trial research team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
- [6] Pinsky PF, Church TR, Izmirlian G, Kramer BS. The National Lung Screening Trial: Results stratified by demographics, smoking history, and lung cancer histology. *Cancer* 2013;119:3976–83.
- [7] Xiang D, Zhang B, Doll D, Shen K, Kloecker G, Freter C. Lung cancer screening: from imaging to biomarker. *Biomark Res* 2013;1:4.
- [8] Adiguzel Y, Kulah H. Biosensors and Bioelectronics Breath sensors for lung cancer diagnosis. *Biosens Bioelectron* 2015;65:121–38.
- [9] Wortley S, Wong G, Kieu A, Howard K. Assessing stated preferences for colorectal cancer screening: a critical systematic review of discrete choice experiments. *Patient* 2014;7:271–82.

- [10] Van Den Bergh KAM, Essink-Bot ML, Van Klaveren RJ, De Koning HJ. Informed participation in a randomised controlled trial of computed tomography screening for lung cancer. *Eur Respir J* 2009;34:711–20.
- [11] Montes U, Seijo LM, Campo a., Alcaide a. B, Bastarrika G, Zulueta JJ. Factors determining early adherence to a lung cancer screening protocol. *Eur Respir J* 2007;30:532–7.
- [12] Wildstein KA, Faustini Y, Yip R, Henschke CI, Ostroff JS. Longitudinal predictors of adherence to annual follow-up in a lung cancer screening programme. *J Med Screen* 2011;18:154–9.
- [13] Ford ME, Havstad SL, Flickinger L, Johnson CC. Examining the effects of false positive lung cancer screening results on subsequent lung cancer screening adherence. *Cancer Epidemiol Biomarkers Prev* 2003;12:28–33.
- [14] Van den Bergh KAM, Essink-Bot ML, Borsboom GJJM, Scholten ET, Prokop M, de Koning HJ, et al. Short-term health-related quality of life consequences in a lung cancer CT screening trial (NELSON). *Br J Cancer* 2010;102:27–34.
- [15] Weernink MGM, Janus SIM, van Til J a., Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceut Med* 2014.
- [16] Phillips KA, Van Bebbler S, Marshall D, Walsh J, Thabane L. A review of studies examining stated preferences for cancer screening. *Prev Chronic Dis* 2006;3:A75.
- [17] Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl Health Econ Health Policy* 2013;11:319–29.
- [18] Mansfield C, Tangka FKL, Ekwueme DU, Smith JL, Guy GP, Li C, et al. Stated Preference for Cancer Screening: A Systematic Review of the Literature, 1990-2013. *Prev Chronic Dis* 2016;13:E27.
- [19] Hummel JM, Steuten LGM, Groothuis-Oudshoorn CJM, Mulder N, IJzerman MJ. Preferences for colorectal cancer screening techniques and intention to attend: A multi-criteria decision analysis. *Appl Health Econ Health Policy* 2013;11:499–507.
- [20] Groothuis-Oudshoorn CGM, Fermont J, IJzerman MJ, Fermont JM, van Til JA, et al. Public stated preferences and predicted uptake for genome-based colorectal cancer screening. *BMC Med Inform Decis Mak* 2014;14:18.
- [21] Ghanouni A, Smith S, Halligan S, Plumb A, Boone D, Yao G, et al. Public preferences for colorectal cancer screening tests: a review of conjoint analysis studies. *Expert Rev Med Devices* 2013;10:489–99.
- [22] Marshall D, McGregor E, Currie G. Measuring Preferences for Colorectal Cancer Screening. *Patient Patient-Centered Outcomes Res* 2010;3:79–89. doi:10.2165/11532250-000000000-00000.
- [23] De Bekker-Grob EW, Rose JM, Donkers B, Essink-Bot M-L, Bangma CH, Steyerberg EW. Men's preferences for prostate cancer screening: a discrete choice experiment. *Br J Cancer* 2013;108:533–41.
- [24] K H, Salkeld G, Patel M, Mann G, Pignone M. Men's preferences and trade-offs for prostate cancer screening: a discrete choice experiment. *Heal Expect* 2014.
- [25] Kaltoft M, Turner R, Nielsen J, Cunich M, Salkeld G, Dowie J. Addressing preference heterogeneity in public policy by combining Cluster Analysis and Multi-Criteria Decision Analysis. *Health Econ Rev* 2014;5:1–11.
- [26] Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. *Med Decis Mak* 2009;27:112–27. doi:10.1177/0272989X06297393.
- [27] Rogowski W, Payne K, Schnell-Inderst P, Manca A, Rochau U, Jahn B, et al. Concepts of “personalization” in personalized medicine: implications for economic evaluation. *Pharmacoeconomics* 2015;33:49–59.

- [28] Eleveld H. Ionising radiation exposure in the Netherlands. Bilthoven: 2003.
- [29] Horeweg N, Scholten ET, de Jong P a., van der Aalst CM, Weenink C, Lammers JWJ, et al. Detection of lung cancer through low-dose CT screening (NELSON): A prespecified analysis of screening test performance and interval cancers. *Lancet Oncol* 2014;15:1341–50.
- [30] Krilaviciute A, Heiss J, Leja M, Kupcinskas J, H H, Brenner H. Detection of cancer through exhaled breath: a systematic review. *Oncotarget* 2015;6:38643–57.
- [31] Hamilton G, Rath B. Detection of circulating tumor cells in non-small cell lung cancer. *J Thorac Dis* 2016;8:1024–8.
- [32] Keeney R, Raiffa H. Decisions with multiple objectives. Cambridge: Cambridge University Press; 1976.
- [33] Belton V, Stewart TJ. Multiple criteria decision analysis : An integrated approach. 2nd ed. Dordrecht: Kluwer Academic Publishers; 2002.
- [34] Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines — part 1 : the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiol Drug Saf* 2007;16:S2–15.
- [35] Felli JC, Noel RA, Cavazzoni PA. A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. *Med Decis Mak* 2009;29:104–15.
- [36] Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;98:270–6.
- [37] Choo EU, Schoner B, Wedley WC. Interpretation of criteria weights in multicriteria decision making. *Comput Ind Eng* 1999;37:527–41.
- [38] Duda RO, Hart PE, Stork DG. Pattern Classification. JOHN WILEY & SONS, INC.; 2000.
- [39] Mooi E, Sarstedt M. Cluster Analysis. A Concise Guid. to Mark. Res. Anal. Process. data, methods using IBM SPSS Stat., 2011, p. 237–84.
- [40] Dunn JC. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J Cybern* 1973;3:32–57.
- [41] Puggina A, Broumas A, Ricciardi W, Boccia S. Cost-effectiveness of screening for lung cancer with low-dose computed tomography: a systematic literature review. *Eur J Public Health* 2015;5:1–8.
- [42] Xie X, Heuvelmans M a, van Ooijen PM a, Oudkerk M, Vliegenthart R. A practical approach to radiological evaluation of CT lung cancer screening examinations. *Cancer Imaging* 2013;13:391–9.

Chapter 7

Modeling multivariate preference distributions using
copula functions

Henk Broekhuizen

Catharina G.M. Groothuis-Oudshoorn

Maarten J. Ijzerman

Andrea Manca

Abstract

Background: Preference elicitation studies have been proposed together with multi-criteria decision analysis (MCDA) models for value-based and patient-centered healthcare decisions. Until now, correlations between attribute weights have been largely ignored. Copulas are a flexible method for fitting multivariate distributions and this study aims to explore their use for fitting multivariate preference distributions.

Methods: We fitted copulas to public preference data for seven attributes of lung cancer screening. Maximum likelihood estimation was used to fit marginal density distributions for each attribute. A vine copula was used to capture dependencies. We compared Monte Carlo samples from the copula to samples obtained with normal, beta, and Dirichlet distributions.

Results: Most fitting methods estimated mean attribute weights and standard deviations well (mean differences ≈ 0). Copulas had the lowest differences in skewness and kurtosis compared to the dataset (0.6 and 4.5, respectively), while the normal distributions had the highest difference (2.3 and 8.4, respectively). The multivariate normal distribution had the highest log likelihood (17843) and the Dirichlet distribution had the lowest (809). The multivariate normal distribution had a lower Frobenius norm than the copula (0.0 compared to 0.1). The MCDA assumptions were met best by the Dirichlet distribution (100%) while the normal distributions yielded negative weights in 23% of samples

Conclusions: Using a large dataset, we have illustrated that copula functions are a flexible and appropriate approach for fitting multivariate preference distributions. We recommend the reporting of copula parameters in preference studies and recommend further research into the application of copulas in personalized healthcare.

Introduction

Even when supported by the best evidence, healthcare policy decisions inevitably require subjective value judgments [1,2]. These need to be as transparent and consistent as possible, and be made using a consistent conceptualization of what constitutes ‘value’ [3,4]. Stakeholders are increasingly asked to inform these value judgments [5,6], in particular patients when alternative decision options may have a profound impact on their prognoses. To take into account patient views in a consistent and representative manner, it has been proposed to construct a quantitative value framework that incorporates patient preferences, elicited with methods such as discrete choice experiments (DCE) or multi-criteria decision analysis (MCDA) [2,7,8]. These methods can be used to describe treatment alternatives in terms of their *attributes*, for example “percentage symptom reduction”, “mode of administration”, or “probability of side effect”. From the patients’ responses to structured questionnaires it is possible to estimate the relative importance weights of these attributes, which shed light on the value that respondents place on specific characteristics of the interventions and the extent to which they are prepared to receive an improvement in one attribute at the cost of a deterioration in another. These attribute weights can for instance be presented to decision makers [8] or be used to explicitly let patients weight the clinical performance of treatments [9].

As patients assign a preference weight to each of the included treatment attributes in preference studies, data from such studies is inevitably multivariate in nature and can be represented as a vector of attribute weights per patient respondent [10]. There can be substantial differences in the (vector of) attribute weights that patients assign to treatment attributes [11]. Acknowledging these differences is a novel direction for the field of personalized healthcare [12]. Nevertheless, the multivariate nature of preferential data has been largely ignored in decision models by assuming that the distributions for attribute weights are independent [9]. This is, however, not a realistic assumption as there can be clear correlations between attribute weights. For example, risk averse patients are more likely to give relatively high weights to all side effects of treatment, which makes it likely that these attribute weights are correlated. Disregarding such correlations between attribute weights may misrepresent the actual underlying preferences of the respondent sample and may subsequently bias probabilistic models that use the preference data as input to construct patient-weighted treatment values. A first method to address the unrealistic assumption of independence between attribute weights is to assume they can be described with a multivariate normal distribution [13].

However, a disadvantage of this approach is that a series of univariate normal distributions will be imposed on the individual attributes. This can be disadvantageous because different types of distributions may provide the best fit for different attributes, depending on the specifics of the attribute. One way to improve on this would be to fit multivariate preference distributions using copulas [14]. Copulas allow for the modeling of the marginal density distributions and dependency relations in two independent but simultaneous steps [15]. Marginal distributions of the attribute weights can be selected using conventional fitting methods for univariate distributions, while the dependencies are parameterized with a copula function. A copula multivariate distribution can therefore contain any type of marginal distribution (normal, beta, uniform, etc.), while still retaining the correlation structure between the various marginal distributions [16].

There is a rich history of applying copulas, primarily in the fields of economics and finance [14,17,18]. There are several applications in healthcare where copulas have been used to model between-location variability in multinational clinical trials outcomes data [19,20]. Another area of application in healthcare can be found in the meta-analysis of diagnostic accuracies literature, where the correlation between sensitivity and specificity prohibits the use of conventional techniques [21]. Despite its apparent advantages, no applications of copulas to model dependency relations in multivariate preference data are known. Therefore, the aim of this paper is to explore the use of copulas for fitting multivariate patient preference data and to compare copulas to other fitting methods using data from a study on public preferences for attributes of lung cancer screening.

Methods

Overview of copulas

A copula function is a multivariate distribution function with uniformly distributed marginal distributions. The word ‘copula’ derives from the Latin word for ‘link’, which is an aptly chosen name as Abe Sklar showed that one can uniquely write a joint distribution function for n random variables in terms of the marginal distribution functions F_i for each variable and a copula function C [22]:

$$F(w_1, \dots, w_n) = C(F_1(w_1), \dots, F_n(w_n)). \quad (1)$$

One can model a multivariate distribution function by estimating the marginals and the copula function separately in two distinct steps: 1) selecting marginal distribution functions

and 2) selecting a copula function. The main advantage of using copulas for modeling multivariate data is that in the first step any type of marginal distribution function (normal, beta, etc.) can be selected as long as its distribution function has an inverse. The reason for this is that the inverse of the distribution function, constructed using the quantile transform, maps all values of $[0,1]$ to w_k . This implies that it can be used as a marginal distribution function for a copula function that has, by definition, uniform marginals.

Most research has focused on selecting and fitting bivariate copula functions [14,23]. Literature abounds about the various ‘families’ of copula functions that describe different types of dependencies [15]. The simple Clayton copula function, for example, defines dependence between two uniformly distributed variables u_1 and u_2 using a single parameter θ :

$$C_{clayton}(u_1, u_2; \theta) = \max\left\{(u_1^{-\theta} + u_2^{-\theta})^{1/\theta}, 0\right\} \quad (2)$$

with $\theta \in (-1, \infty)$ and $\theta \neq 0$. Using this copula, one can represent a negative correlation with $\theta = -1$, independence with $\theta = 0$, and a positive correlation with θ greater than 1 [24].

Practically fitting copulas in higher dimensions

In the case of preference data, we are interested in modeling dependency between a larger number of attributes. The available (families of) copula functions that can deal with this in a practical manner is limited [25]. Previous studies found that a practical approach for constructing a higher dimensional copula function is to pairwise combine bivariate copula functions in several steps [16,26]. Copula functions constructed in this way are termed *vine copulas*. We will illustrate the rationale behind vine copulas for the most simple case with three random variables W_1 , W_2 and W_3 . See that by applying the chain rule the joint density function of these variables f can be written as a product of conditional density functions and one marginal density function, e.g.

$$f(w_1, w_2, w_3) = f(w_3|w_2, w_1)f(w_2|w_1)f(w_1) \quad (3)$$

Furthermore, by differentiating Formula 1 one can write a joint density function as a product of a copula density function and marginal density functions. For $n = 2$ for example the joint density function is $f(w_1, w_2) = c_{12}f_1f_2$ where c_{12} is shorthand for the copula density $c(F_1(w_1), F_2(w_2))$ and f_k is shorthand for the density $f(w_k)$. With this in mind one

can rewrite the conditional density of, for example, W_2 as a function of a copula density function and the marginal density function of W_2 :

$$f(w_2|w_1) = \frac{f_{1,2}}{f_1} = c_{12}f_2 \quad (4)$$

This illustrates that we can express the joint density function as a product of conditional density functions and a marginal density function (Formula 3) and we can express conditional density functions as a product of bivariate copula density functions and marginal density functions (Formula 4). This implies that we can write the joint density function f for the three variables in our example as a product of bivariate copula density functions and marginal density functions:

$$f(w_1, w_2, w_3) = c_{23|1}c_{12}c_{13}f_1f_2f_3, \quad (5)$$

With, c_k shorthand notation for the copula density, e.g. $c_{23|1} = c(F_2(w_2), F_3(w_3)|F_1(w_1))$. A joint density function on n dimensions can be represented as a series of $n - 1$ undirected graphs ('trees'), where the edges represent bivariate copula functions and where the edges of graph j become the nodes of tree $j + 1$ [27,28]. An example of the vine copula described by Formula 5 is given in Figure 1. Importantly, this vine copula structure is not unique as in fact there are three alternative structures. The number of possible vine copula structures increases quadratically with n , more precisely with $n(n - 1)/2$ [24]. The choice of vine copula structure is important as some structures have a better fit to the data than others. There are multiple ways to select a vine copula structure that best preserves the dependency relations in the data [16]. In general, for a better fit of the data under study it has been found to be advantageous to combine pairs of attributes with a higher correlation before combining pairs with a lower correlation [24].

Figure 1: Vine copula structure of the simple three variable example. The copula function of Formula 5 can be visualized using this vine copula. Each node in the first tree represents a random variable, and each edge represents a copula function. Nodes in tree $j, j > 1$ represent the edges of tree $j - 1$.

Applying copulas for fitting multivariate preference data

In this study, copulas will be applied for the case of lung cancer screening. Lung cancer is among the most deadly types of cancer in the world, especially so because symptoms usually present only at later stages [29]. Since the survival for early stage cancer is much better than that for later stages, there is increasing attention for developing lung cancer screening programs [30,31]. However, it is unknown how one should operationalize these in such a way that attendance is maximized [32]. Therefore, it is important to know what factors determine whether or not someone decides to attend a screening. In an earlier study we elicited preferences of the general population that may be eligible for screening in the Netherlands [33]. During interviews and a panel session the following attributes of screening were identified: sensitivity, specificity, radiation load, duration of screening procedure, waiting time until results, mode of screening, and location of screening. Subsequently, a questionnaire was used to assess the relative weights of these attributes according to a sample of the Dutch general population. A total of 1034 complete responses to the questionnaire were obtained, which contained for each respondent a vector of seven attribute weights, i.e. one for every attribute. By design, each weight is between zero and one, and the attribute weights add up to one per respondent. It is the multivariate distribution of the attribute weights that we aim to model with the use of copulas. To do so, we have to select the best fitting probability density function for each attribute and select an appropriate structure for the vine copula.

Selecting marginal distributions

Since not all marginal density functions need to be of the same type when using copulas, a choice of marginal density function has to be made for each attribute. For this we used maximum likelihood estimation to fit several types of distributions for each attribute, and we then selected the best fitting distribution based on minimizing the Bayesian information criterion (BIC) [34]. As candidates we considered the following distributions. First of all, the Beta distribution was included as it is naturally defined (as the attribute weights are) on the $[0,1]$ interval. For similar reasons we also included the Gamma, Weibull, F, and exponential distributions because they are defined on the $[0,\infty]$ interval. Third of all we included the normal and Student t distributions as they are very commonly applied distributions. Finally, because descriptive statistics and empirical density plots of the data showed signs of bimodality in the weights for some attributes, we also included bimodal beta mixture distributions [35]. Since such mixture distributions are not commonly applied in the field of preferences studies and because we suspected that they would dominate the simpler options described above, we opted to use them only as marginal distributions for a separate copula-based multivariate distribution. For optimization purposes, attribute

weights in the data of exactly 0.00 were altered to 10^{-6} and attribute weights of exactly 1.00 were altered to $1 - 10^{-6}$.

Selecting the vine copula structure

The structure for the vine copula was found using the data-driven method based on Kendall's τ described in [26]. The method consisted of six steps. In each step a tree was selected that maximized the pairwise correlations as measured with Kendall's τ [36]. This process was continued until a complete vine copula function was obtained for the whole dataset. For each pair in the vine, the best fitting type of copula function was selected based on the BIC. As possible options we considered the following (families of) copula functions: the independence copula, the Gaussian copula, the Student t copula, the Clayton copula, the Gumbel copula, the Frank copula, the Joe copula, the BB1 copula, the BB6 copula, the BB7 copula, and the BB8 copula [24]. Rotated versions of these copula functions were also included to be able to fit negative dependencies.

Simulation study

To study the performance of the fitting of multivariate preference distributions with vine copulas, we compared their fit to the fit of other commonly used multivariate distributions. This was done as follows. First of all, we selected several commonly used distributions as comparators and estimated their parameters using maximum likelihood estimation. Then, we drew samples from both the copula distribution and the comparators. These samples were then compared to the original data using a set of predefined comparison metrics.

Multivariate distributions used as comparators

We started by fitting two multivariate distributions to our data, namely the multivariate normal distribution and the Dirichlet distribution. The first was used in an earlier study that modelled correlations between attribute weights [13], and the latter is the multivariate extension of the Beta distribution and seems especially relevant for our case of preference data since it, like the attribute weights, is defined on $[0,1]^n$. Since many previous decision modeling studies have assumed independence between attribute weights, we also fitted univariate beta and univariate normal distributions to the preference data. Parameters for all of these distributions were obtained with maximum likelihood estimation.

Sampling scheme

In total, 10,000 samples were generated for each method. Each sample consisted of 1034 observations to make the sample size equal to that of the original dataset. To obtain a sample from the copula-based multivariate distributions, first a sample from the hypercube

$[0,1]^n$ was drawn from the copula function [27]. By using the inverse distribution function of the fitted marginal distribution functions these random vectors could be transformed into a sample of attribute weights. Samples for the other methods were obtained by inserting samples from a uniformly distributed hypercube $[0,1]^n$ into the inverse (i.e. quantile) distribution functions of each comparator, references for which can be found in [37–40].

Comparison metrics

We compared the samples from the six different methods with seven performance metrics. There are five per-attribute metrics that describe how well the characteristics of individual attribute are captured, and two overall metrics that describe how well the dataset as a whole is captured. The per-attribute metrics are differences between the data and the samples with regard to their mean, standard deviation, skewness, and kurtosis. The first overall metrics was the Frobenius norm (i.e. the generalization of the Euclidean distance) between the covariance matrix of a sample and the covariance matrix of the original dataset. Here a low Frobenius norm implies that the covariance matrix of a sample has a good resemblance to the covariance matrix of the dataset. For each fitted distribution we also calculated the log likelihood and we estimated the percentage of attribute weights

	Sensitivity	Specificity	Radiation load	Duration of procedure	Wait until results	Mode	Location
<i>beta</i>	-1739	-2169	-1961	-2658	-2155	-1631	-1467
<i>exp</i>	-1716	-2124	-2091	-2666	-2160	-1622	-1494
<i>f</i>	-516	-868	-652	-1231	-803	-370	-409
<i>gamma</i>	-1715	-2138	-2085	-2677	-2157	-1615	-1500
<i>logistic</i>	-1256	-1653	-1734	-2370	-1804	-1246	-943
<i>normal</i>	-1087	-1473	-1417	-1942	-1464	-1066	-773
<i>t</i>	-1350	-1718	-1882	-2511	-1979	-1373	-1112
<i>weibull</i>	-1709	-2119	-2095	-2659	-2154	-1624	-1488
<i>Bimodal mixture beta</i>	-1992	-2455	-2407	-3231	-2604	-1886	-1757

Table 1: BIC values for the marginals for the included parametric univariate distributions.

that were between zero and one across all samples obtained from a fitting method. Finally, we calculated the percentage of samples where all attribute weights summed up to one. For this metric we allowed for an error of 0.01. All of the above metrics (apart from the log likelihood) were calculated for each of the 10,000 obtained samples. Therefore, we can assess their posterior distributions and we chose to use the posterior means of the comparison metrics as the main outcomes for the simulation study.

Results

Copula fitting

The BIC values for the included distribution types found in the first step of fitting a copula-based distribution can be found in Table 1. The beta mixture distribution had the lowest BIC across all attributes. The best fitting other distribution for the *sensitivity*, *specificity*, and *mode of screening* attributes was the beta distribution. For *duration* of procedure and *location* of screening this was a gamma distribution, for *radiation load* this was a Weibull distribution and for *time until results* this was the exponential distribution.

The parameter estimates for these distributions can be found in Table 2. The parameters for the beta mixture distributions are presented in Table 3. The parameter estimates together with standard errors for the univariate normal, multivariate normal, beta, and Dirichlet distributions (i.e. the comparators) can be found in Table 4.

The structure of the vine copula can be found in Figure 2. For the first order pairs (i.e. c_{ij}) Joe-Gumbel, t, and Frank copulas were found to have the lowest BIC. For all the second order pairs (i.e. $c_{ij|k}$), the Student t copula had the lowest BIC. In the third until the sixth order pairs some pairs were best fitted with a (skewed) survival copula. For example, a survival Joe-Frank copula was found to be the best fitting copula for joining *radiation* and *mode of screening*, conditional on *sensitivity* and *specificity*.

Attributes that were combined early on in the process of constructing the vine copula were often aspects of screening that have a clear and logical relation. An example of this is the sensitivity/specificity pair because both attributes describe diagnostic accuracy, i.e. the core clinical benefit of screening. Another example is the duration/wait pair because these attributes are both time related.

In Table 5 we present a complete overview of all fitted bivariate copula functions that were used to construct the vine copula.

Attribute	Type	Parameter 1 (SE)	Parameter 2 (SE)
<i>Sensitivity</i>	Beta	0.80 (0.03)	4.18 (0.20)
<i>Specificity</i>	Beta	0.76 (0.03)	5.10 (0.25)
<i>Radiation load</i>	Weibull	1.08 (0.03)	0.14 (0.00)
<i>Duration of procedure</i>	Gamma	0.85 (0.03)	8.46 (0.43)
<i>Time until results</i>	Exponential	7.75 (0.24)	-
<i>Mode of screening</i>	Beta	0.87 (0.03)	4.29 (0.20)
<i>Location of screening</i>	Gamma	0.87 (0.03)	4.90 (0.25)

Table 2: Types and parameters of best fitting marginals for use in copula distribution

Attribute	Parameters and weight 1st beta (α, β, w)	Parameters and weight 2nd beta (α, β, w)
<i>Sensitivity</i>	[0.16, 0.73, 0.08]	[1.4, 7.70, 0.92]
<i>Specificity</i>	[1.45, 9.67, 0.91]	[0.15, 0.98, 0.08]
<i>Radiation load</i>	[2.16, 16.26, 0.87]	[0.24, 0.79, 0.13]
<i>Duration of procedure</i>	[0.16, 0.99, 0.11]	[2.31, 21.96, 0.89]
<i>Time until results</i>	[2.44, 19.12, 0.85]	[0.23, 1.04, 0.15]
<i>Mode of screening</i>	[0.42, 1.52, 0.31]	[2.89, 18.38, 0.69]
<i>Location of screening</i>	[0.39, 1.19, 0.36]	[2.73, 18.10, 0.64]

Table 3: Parameters of best fitting beta mixture distributions for use in copula distribution

Attribute	Normal		Beta		Dirichlet
	Mean	SD	α_1	α_2	$\alpha_k / \sum \alpha$
<i>Sensitivity</i>	0.16 (0.00)	0.14 (0.00)	0.79 (0.03)	4.20 (0.20)	0.15
<i>Specificity</i>	0.13 (0.00)	0.18 (0.00)	0.76 (0.03)	5.10 (0.25)	0.13
<i>Radiation load</i>	0.13 (0.00)	0.12 (0.00)	0.83 (0.03)	5.00 (0.24)	0.14
<i>Duration of procedure</i>	0.10 (0.00)	0.09 (0.00)	0.78 (0.03)	6.73 (0.33)	0.12
<i>Time until results</i>	0.13 (0.00)	0.12 (0.00)	0.82 (0.03)	5.46 (0.27)	0.14
<i>Mode of screening</i>	0.17 (0.00)	0.14 (0.00)	0.87 (0.03)	4.29 (0.20)	0.16
<i>Location of screening</i>	0.18 (0.01)	0.17 (0.00)	0.71 (0.03)	3.10 (0.15)	0.16

Table 4: Parameter estimates (with standard error) for conventional fitting methods.

Comparison with other methods

Figure 3 shows the marginal density functions each of the fitting methods overlaid on top of the histograms of the original dataset. From this plot it can be seen that not all methods fit the attribute weight distributions equally well. The normal distributions misrepresent the lower bound of the weights, having density for negative weights and also missing density around the low weights because of a lack of skewness. Figure 3 also shows that the marginal density functions for the univariate Beta, Dirichlet, and copula distributions are very similar. The densities for the copula distribution with beta mixture marginal densities generally have a high density around weights of zero, and at 0.1 have a second density peak.

All fitting methods estimated the mean weights and standard deviations well (mean differences all close to zero), with the exception of the Dirichlet distribution. The differences in skewness for the normal distributions are 2.3 on average, 0.6 on average for

Figure 2: An illustration of the sequential building of the Vine copula structure for the lung cancer screening case. Attribute names have been shortened to numbers for clarity, with 1=sensitivity, 2=specificity, 3=radiation load, 4=duration of screening procedure, 5=time until results, 6=mode of screening, 7=location of screening. For details on what copula functions (and associated parameters) were used in the various trees, see Table 5.

the copula distributions and 1.1 on average for the copula distribution with mixture beta marginals. The average differences in skewness for the beta and dirichlet distribution were 0.97 and 1.00, respectively.

The kurtosis metric shows a similar picture: the difference is 8.50 on average for the normal distributions, 4.42 for the copula distribution and 7.4 for the copula distribution with beta mixture marginals. The performance on this metric of the other distributions varies greatly per attribute.

The difference in kurtosis for the radiation load and screening duration attributes are noticeably higher across all fitting methods (14.36 on average) compared to the other attributes (4.00 on average).

The average Frobenius norm for the multivariate normal distribution is 0.00 compared to the values of 0.03 for the independently assumed beta distributions, 0.02 for the Dirichlet distribution, 0.10 for the copula distribution and 0.70 for the copula distribution with beta mixture marginals.

The multivariate normal distribution had the highest log likelihood, namely 17843, followed by the copula distribution with beta mixture marginals (15205), copula distribution (13950), independent beta distributions (6939), independent normal distributions (4659), and the Dirichlet distribution (809).

For the univariate beta distributions, the Dirichlet distribution, and the copula distributions all sampled weights are between zero and one (by definition). For the univariate and multivariate normal distributions negative values for attribute weights were drawn in 13% of the weight samples from the (univariate and multivariate) normal distributions across all attributes.

The Dirichlet and multivariate normal distributions were the only distributions that consistently yielded weight samples whose sum was between 1 ± 0.01 . For the other distributions the sum of weights in around 1.5% of samples was equal to 1 ± 0.01 . When the ± 0.01 tolerance in sampling error was removed, the Dirichlet was (as expected) the only distribution that yielded weight samples that exactly added up to one.

An overview of the comparison metrics across all fitting methods can be found in Table 6.

Tree level	Copula pair	Copula family	P1	P2
1	Rad,spec	Joe-Gumbel	1.63	2.28
	Spec, sens	t	0.74	2.20
	Wait, time	Joe-Gumbel	1.36	2.75
	Mode, sens	Frank	5.32	-
	Sens, location	Frank	4.56	-
	Time, location	Joe-Gumbel	1.63	2.03
2	Rad,sens spec	t	0.22	17.39
	Spec, mode sens	t	0.25	17.42
	Wait, location time	t	0.24	5.89
	Mode, location sens	t	0.57	3.21
	Sens, time location	t	0.46	7.25
3	Rad, mode sens, spec	Survival Joe-Frank	1.67	0.97
	Spec, location mode, sens	Survival Joe-Frank	1.58	0.87
	Wait, sens location, time	Survival Gumbel	1.04	-
	Mode, time location, sens	Survival Joe-Frank	1.77	0.89
4	Rad, location mode, sens, spec	t	0.06	20.78
	Spec, time location, mode, sens	Frank	2.99	-
	Wait, mode sens, location, time	Survival Joe-Frank	1.52	0.87
5	Rad, time location, mode, sens, spec	t	0.22	15.61
	Spec, wait time, location, mode, sens	Joe-Frank	1.42	0.75
6	Rad, wait time, location, mode, sens, spec	Survival Joe-Frank	1.13	0.99

Table 5: Copula pairs with used copula functions and parameters. sens=sensitivity, spec=specificity, rad=radiation load, dur=duration of screening, time=waiting time until results, P1=first parameter, P2=second parameter.

Figure 3: histogram of the attribute weights in the original dataset and marginal density functions each of the fitting methods. Since the parameters estimates for the univariate and multivariate normal are the same, their densities are plotted as one 'normal' density. sens=sensitivity, spec=specificity, rad=radiation load, time=duration of screening procedure, wait=time until results, mode=mode of screening, location=location of screening

Discussion

In this study we explored how copulas can be used to model multivariate preference data that was elicited in an MCDA framework. We compared copulas to conventional fitting methods using a simulation study. Conventional methods and the multivariate normal distribution in particular, outperform copula functions on mean, standard deviation, log likelihood and Frobenius norm. However, the flexibility of arbitrary marginals means that the specific functional forms of marginals were better captured with copulas, as reflected in its better performance on the skewness and kurtosis metrics as well as on the metric regarding the MCDA assumption that weights should be between zero and one. The marginal attribute distributions across all fitting methods were sometimes quite different in functional form (as evidenced in the skewness and kurtosis metrics) to the empirical attribute distributions, and this was especially true for the radiation load and duration of screening attributes.

A perceived bimodality in the empirical densities of some attribute weight encouraged us to try and fit a copula model that had bimodal beta mixture distributions as its marginals. These mixture marginals had a lower (i.e. better) BIC than the other univariate distributions, which was most likely due to them having more parameters (five as opposed to one or two for the other distributions). Nevertheless, the copula function with mixture marginals performed worse than the copula function that had several different types of

marginals on the univariate as well as multivariate comparison metrics. This could imply that it might be advantageous to explore other metrics than the Bayesian information criterion to select the type of marginal densities. The exploration of using mixture beta distributions for the marginal densities points toward an advantage of the copula methodology. This advantage is that, although multivariate mixture distributions are not commonly used or described in literature, we could readily create one because we only needed to fit the marginal densities and could subsequently model the dependencies with the (already fitted) copula function. We elected to use the mixture beta marginals for all attributes in the dataset since all empirical distributions showed signs of bimodality. However, since marginal densities are estimated separately from the copula function, for other preference datasets it may be advantageous to use a mixture distribution for only a subset of attributes.

Our results suggest that vine copula distributions can describe attribute weights properly and in a flexible manner. This would argue for an inclusion of copulas in studies describing public or patient preferences. With public and private healthcare decisions being made more and more based on quantitative models that require both patient preferences and clinical evidence as input, copulas can be a useful tool to model multivariate datasets. Copula models can improve the grounding of decisions in evidence as researchers or policy makers would have a better idea of the dependency relations in the evidence compared to when univariate distributions and/or simple measures of dispersion are used. Furthermore, obtaining or providing complete datasets can present researchers with practical and legal barriers. Our study shows that, using copulas, a multivariate distribution can be described using a relatively small number of data points. One might imagine a table or appendix that describes the form and parameters of each marginal density, a description of the vine copula structure, and the copula functions with parameters. Given m patients providing their preferences for n attributes, we could then describe the data with $(n - 1) + (n - 2) + \dots + 1$ bivariate copula functions (each having one or two parameters), an order of combination for the vine construction and n marginal densities (each with one or two parameters), which is much smaller than reporting the whole dataset (which has a size of nm). The reporting of copula structure and parameters may thus present a practical midpoint between incomplete reporting and the practical/legal challenges of providing complete datasets.

Reporting copula functions and parameters in preference studies still has two practical limitations. First of all, fitting copula models requires some knowledge of statistical software (R, SAS or Stata) as it is not a standard option in commonly used packages like

	Independent normal	Independent beta	Multivariate normal
<i>Difference in means</i>	<i>Sens</i>	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)
	<i>Spec</i>	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)
	<i>Rad</i>	0.00 (-0.01 to 0.01)	-0.01 (-0.02 to 0.00)
	<i>Dur</i>	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.00)
	<i>Time</i>	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)
	<i>Mode</i>	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)
	<i>Location</i>	0.00 (-0.01 to 0.01)	-0.01 (-0.02 to 0.00)
<i>Difference in SD</i>	<i>Sens</i>	0.00 (-0.01 to 0.01)	-0.01 (-0.02 to 0.00)
	<i>Spec</i>	0.00 (0.00 to 0.01)	-0.01 (-0.02 to 0.00)
	<i>Rad</i>	0.00 (-0.01 to 0.01)	-0.01 (-0.02 to -0.01)
	<i>Dur</i>	0.00 (0.00 to 0.00)	-0.01 (-0.02 to 0.00)
	<i>Time</i>	0.00 (0.00 to 0.01)	-0.01 (-0.01 to 0.00)
	<i>Mode</i>	0.00 (-0.01 to 0.01)	-0.01 (-0.02 to 0.00)
	<i>Location</i>	0.00 (-0.01 to 0.01)	-0.01 (-0.02 to 0.00)
<i>Difference in skewness</i>	<i>Sens</i>	1.83 (1.69 to 1.99)	0.54 (0.37 to 0.70)
	<i>Spec</i>	1.96 (1.81 to 2.10)	0.50 (0.30 to 0.69)
	<i>Rad</i>	2.64 (2.50 to 2.79)	1.28 (1.10 to 1.46)
	<i>Dur</i>	3.66 (3.51 to 3.81)	2.06 (1.83 to 2.29)
	<i>Time</i>	2.65 (2.50 to 2.80)	1.23 (1.03 to 1.41)
	<i>Mode</i>	1.83 (1.68 to 1.98)	0.61 (0.45 to 0.76)
	<i>Location</i>	1.80 (1.66 to 1.95)	0.59 (0.44 to 0.72)
<i>Difference in kurtosis</i>	<i>Sens</i>	4.19 (3.86 to 4.46)	2.75 (1.97 to 3.41)
	<i>Spec</i>	5.54 (5.21 to 5.81)	3.39 (2.32 to 4.28)
	<i>Rad</i>	10.28 (9.96 to 10.54)	8.51 (7.59 to 9.28)
	<i>Dur</i>	22.21 (21.89 to 22.47)	19.33 (17.8 to 20.50)
	<i>Time</i>	9.62 (9.29 to 9.89)	7.55 (6.45 to 8.42)
	<i>Mode</i>	4.04 (3.71 to 4.31)	2.80 (2.10 to 3.40)
	<i>Location</i>	3.45 (3.13 to 3.72)	2.43 (1.86 to 2.93)
<i>Frobenius</i>	0.03 (0.03 to 0.03)	0.03 (0.03 to 0.03)	0.00 (0.00 to 0.01)
<i>LL</i>	4659	6939	17843
<i>P01</i>	0.87 (0.86 to 0.87)	1.00 (1.00 to 1.00)	0.87 (0.86 to 0.87)
<i>Ps1</i>	0.02 (0.01 to 0.03)	0.02 (0.01 to 0.03)	1.00 (1.00 to 1.00)

Table 6: Comparison metrics with empirical 2.5%th and 97.5%th quantiles. Cov=covariance, sens=sensitivity, spec=specificity, rad=radiation load, dur=duration of screening, time=waiting time until results, LL=log-likelihood, P01=percentage of weights between 0 and 1, PD1=percentage of samples with sum of weights 1 ± 0.01 .

Dirichlet	Copula with marginals of Table 1a	Copula with Beta mixture
0.01 (0.00 to 0.02)	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)
0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)	0.00 (-0.01 to 0.01)
-0.01 (-0.02 to 0.00)	0.00 (-0.01 to 0.01)	0.00 (0.00 to 0.01)
-0.02 (-0.02 to -0.01)	0.00 (-0.01 to 0.01)	0.00 (0.00 to 0.01)
-0.01 (-0.01 to 0.00)	0.00 (-0.01 to 0.01)	0.01 (0.00 to 0.01)
0.01 (0.00 to 0.01)	0.00 (-0.01 to 0.01)	0.01 (0.00 to 0.01)
0.02 (0.01 to 0.03)	0.00 (-0.01 to 0.01)	0.01 (0.00 to 0.01)
0.01 (0.00 to 0.02)	-0.01 (-0.02 to 0.00)	0.02 (0.01 to 0.02)
-0.01 (-0.02 to 0.00)	-0.01 (-0.02 to 0.00)	0.01 (0.01 to 0.02)
-0.01 (-0.02 to 0.00)	0.00 (-0.01 to 0.01)	0.02 (0.02 to 0.03)
-0.02 (-0.03 to -0.02)	-0.02 (-0.03 to -0.01)	0.02 (0.01 to 0.02)
-0.01 (-0.02 to 0.00)	-0.01 (-0.02 to 0.00)	0.02 (0.02 to 0.03)
0.01 (0.00 to 0.01)	-0.01 (-0.02 to 0.00)	0.02 (0.01 to 0.03)
0.03 (0.02 to 0.04)	-0.02 (-0.04 to -0.01)	0.02 (0.02 to 0.03)
0.58 (0.40 to 0.74)	0.54 (0.38 to 0.7)	0.64 (0.49 to 0.79)
0.55 (0.35 to 0.73)	0.50 (0.31 to 0.69)	0.61 (0.42 to 0.78)
1.32 (1.13 to 1.49)	0.89 (0.43 to 1.22)	1.54 (1.40 to 1.67)
2.11 (1.89 to 2.31)	1.53 (0.90 to 1.95)	2.23 (2.05 to 2.39)
1.27 (1.08 to 1.45)	0.68 (0.12 to 1.07)	1.4 (1.25 to 1.54)
0.64 (0.47 to 0.81)	0.61 (0.45 to 0.76)	0.74 (0.61 to 0.87)
0.59 (0.42 to 0.76)	-0.31 (-0.93 to 0.10)	0.78 (0.66 to 0.90)
2.72 (1.84 to 3.43)	2.75 (2.00 to 3.39)	3.08 (2.38 to 3.67)
3.53 (2.47 to 4.38)	3.40 (2.36 to 4.28)	3.76 (2.80 to 4.52)
8.58 (7.64 to 9.33)	5.86 (1.51 to 8.07)	9.66 (9.08 to 10.12)
19.65 (18.36 to 20.65)	15.51 (8.49 to 18.9)	20.28 (19.35 to 21.01)
7.70 (6.70 to 8.54)	3.88 (-2.05 to 6.79)	8.50 (7.85 to 9.05)
2.79 (2.02 to 3.45)	2.81 (2.10 to 3.40)	3.48 (3.00 to 3.89)
2.13 (1.32 to 2.81)	-3.09 (-9.78 to 0.13)	3.26 (2.88 to 3.60)
0.02 (0.02 to 0.02)	0.10 (0.09 to 0.12)	0.07 (0.07 to 0.08)
809	13950	15205
1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
1.00 (1.00 to 1.00)	0.01 (0.00 to 0.01)	0.01 (0.00 to 0.02)

Table 6 (continued from previous page)

SPSS and Excel. We chose to work in R because several authors have released copula-related packages and because it is free (and therefore more approachable than commercial packages). Furthermore, we have provided the R code that underlies the analyses in this paper as supplementary material to the published paper. As a second practical limitation, it can be hard to interpret the meaning of copula function parameters, especially in problems with much more than two dimensions. For example, that the parameter θ in $C_{spec,time|location,mode,sens} = Frank(\theta)$ equals 2.99 does not at first glance confer insight into the data or the dependency relations, despite its good fit. This could lead to copulas being perceived as a black box by healthcare decision makers. Care should thus be taken that the complexity of copulas does not undermine the trust decision makers place in the data or the decision models.

The main strength of our study is that we were able to compare a variety of fitting methods on a large preference dataset. Furthermore, by drawing a large number of samples, we were able to investigate the posterior distributions of seven comparison metrics. Previous MCDA studies that use multivariate preference data whilst also taking into account dependency relations mostly use multivariate normal distributions [13]. Although this is an improvement compared to assuming independence [9], this approach has several drawbacks.

First of all, the marginal density functions of the attribute weights have to take the same parametric (in the case of the normal distribution: symmetrical) functional form, which restricts the flexibility of the approach when the number of attributes increases. This is clearly reflected in the low scores on the skewness and kurtosis metrics for some attributes in our study. Secondly, the multivariate normal distribution uses only one parameter to describe bivariate dependency relations (namely, a particular element in the covariance matrix). This precludes practitioners from modeling higher order dependency relations enabled by parameterizing dependence with more than one parameter. Finally, samples drawn from normal distributions can fall outside the (0,1) support, thus violating MCDA assumptions. The single study we could find on the use of copula for MCDA was a 2014 conference paper on finding probability distributions for a small set of performance scores [41], with the aim to generate a larger performance dataset for testing purposes. Our study has an advantage over that study because we were able to compare the samples from the fitted distributions with a large dataset.

Our study has some limitations. First of all, it is possible that our comparison framework excluded important metrics. We used the Frobenius norm between the covariance matrix

of the original data and the covariance matrix of a randomly drawn sample from fitted distributions as a measure of dependency fit. Since covariance matrices capture bivariate dependencies only, and with only one parameter, they may present us with a too simple view on the extent of dependency fit. The multivariate normal distribution had the smallest Frobenius norm, but this is logical since the same matrix was used as its parameter estimator and we did not quantify the extent to which other or higher order dependency relations were captured. It would be possible to include all (differences in) covariances as separate metrics, but this would make the study results harder to interpret since it would add n^2 (i.e. 49) metrics. A third limitation is that our simulation study considered only one particular preference sample in one particular health area. As such, our results and recommendations cannot be generalized to all preference studies. Finally, the vine copula structure we selected was chosen based on BIC, but it is possible that using another optimization parameter would have led to another structure and thus to a different performance compared to the conventional fitting methods.

In this study we fitted the distribution for the complete public sample. However, the preference study we used the data from also identified preference clusters. It would be interesting to fit distributions for the preferences of these clusters. It may be possible that the aforementioned issue with multimodal weight distributions is solved when the dataset is split up into clusters. In the present study we had a large dataset to work with while in some situations only smaller preference datasets may be available. It would be useful to know which methods are best for fitting distributions on smaller preference datasets. One way to get an idea for this would be to repeat the simulation study in this paper for smaller subsets of the dataset, e.g. repeated fitting on a random subset of 10%, 20%, ..., and 90% of the data and comparing how quickly each of the methods' metrics converge to the values found when 100% of the data is fitted, and the extent to which estimates vary on smaller datasets. One might also imagine another approach where the distribution is fitted on a *training set* that comprises 70% of the data and that the comparison metrics are then calculated based on the remaining 30% of the data (the *validation set*).

Finally, a promising research direction for personalized healthcare may be the joint modeling of preference heterogeneity and performance heterogeneity (e.g. a drug being more efficacious for a particular age group). Copulas may be particularly well suited for this type of study as the different scales that clinical outcomes are measured on would not pose a problem for copula modeling. Instead, best-fitting distributions of an arbitrary type can be found for each attribute weight and clinical outcome, while the copula function can be fitted separately to model the dependency relations.

In conclusion, copula functions are a flexible and appropriate approach for fitting multivariate preference distributions. The main advantage over conventional approaches is that marginal attribute density functions and dependencies can be modeled separately. This enables practitioners investigating correlations in their models to choose those marginal density functions that best fit the data instead of being required to use the marginal density function that belongs to the prior choice of multivariate distribution.

References

- [1] Hofmann B, Cleemput I, Bond K, Krones T, Droste S, Sacchini D, et al. Revealing and Acknowledging Value Judgments in Health Technology Assessment. *Int J Technol Assess Health Care* 2015;30:579–86.
- [2] Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl Health Econ Health Policy* 2013;11:319–29.
- [3] Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. *Value Heal* 2010;13:657–66.
- [4] Yuan Z, Levitan B, Berlin JA. Benefit-risk assessment: to quantify or not to quantify, that is the question. *Pharmacoepidemiol Drug Saf* 2011;20:653–6.
- [5] Van Til JA, IJzerman MJ. Why Should Regulators Consider Using Patient Preferences in Benefit-risk Assessment? *Pharmacoeconomics* 2013;10–3.
- [6] Bridges JFP, Jones C. Patient-based health technology assessment: a vision of the future. *Int J Technol Assess Health Care* 2007;23:30–5.
- [7] Weernink MGM, Janus SIM, van Til JA., Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceut Med* 2014.
- [8] Oude Egbrink M, IJzerman MJ. The value of quantitative patient preferences in regulatory benefit-risk assessment. *J Mark Access Heal Policy* 2014;2:1–4.
- [9] Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber a B, Jansen JP, IJzerman MJ. Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis. *BMC Med Inform Decis Mak* 2015;15:1–10.
- [10] Broekhuizen H, Groothuis-Oudshoorn CGM, van Til JA, Hummel JM, IJzerman MJ. A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics* 2015;33:445–55.
- [11] Hauber AB, Mohamed AF, Watson ME, Johnson FR, Hernandez JE. Benefits, risk, and uncertainty: preferences of antiretroviral-naïve african americans for HIV treatments. *AIDS Patient Care STDS* 2009;23:1–6.
- [12] Rogowski W, Payne K, Schnell-Inderst P, Manca A, Rochau U, Jahn B, et al. Concepts of “personalization” in personalized medicine: implications for economic evaluation. *Pharmacoeconomics* 2015;33:49–59.
- [13] Broekhuizen H, IJzerman MJ, Hauber AB, Groothuis-oudshoorn CGM. Weighing clinical evidence using patient preferences: an application of probabilistic multi-criteria decision analysis. *Pharmacoeconomics* 2016.

- [14] Genest C, Gendron M, Bourdeau-Brien M. The advent of copulas in finance. *Eur J Financ* 2009;15:609–18.
- [15] Rüschendorf L. Copulas, Sklar's theorem, and distributional transform. *Math. risk Anal.*, Berlin Heidelberg: Springer; 2013, p. 3–33.
- [16] Bedford T, Daneshkhah A, Wilsin KJ. Approximate uncertainty modeling in risk analysis with vine copulas. *Risk Anal* 2016;36:792–815.
- [17] Nelsen R. Copulas, Characterization, Correlation, and Counterexamples. *Math Mag* 1995;68:193–8.
- [18] Nelsen R. An Introduction to Copulas. New York, NY: Springer New York; 2006.
- [19] Manca A, Lambert PC, Sculpher M, Rice N. Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. *Med Decis Mak* 2007;27:471–90.
- [20] Nixon R, Thompson S. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ* 2005;14:1217–29.
- [21] Kuss O, Hoyer A, Solms A. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Stat Med* 2014;33:17–30.
- [22] Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publ L'institute Stat L'université Paris* 1959;8:229–31.
- [23] Kolev N, Dos Anjos U, Mendes B. Copulas: A review and recent developments. *Stoch Model* 2006;22:617–60.
- [24] Aas K, Czado C, Frigessi A, Bakken H. Pair-copula constructions of multiple dependence. *Insur Math Econ* 2009;44:182–98.
- [25] Kurowicka D, Cooke R. Uncertainty Analysis with High Dimensional Dependence Modelling. Chichester, UK: John Wiley & Sons, Ltd; 2006.
- [26] Brechmann EC, Czado C, Aas K. Truncated regular vines in high dimensions with applications to financial data. *Can J Stat* 2012;40:68–85.
- [27] Dissmann JF, Brechmann EC, Czado C, Kurowicka D. Selecting and estimating regular vine copulae and application to financial returns. *Comput Stat Data Anal* 2013;59:52–69.
- [28] Gruber L, Czado C. Sequential bayesian model selection of regular vine copulas. *Bayesian Anal* 2015;10:937–63.
- [29] Ades AE, Biswas M, Welton NJ, Hamilton W. Symptom lead time distribution in lung cancer: natural history and prospects for early diagnosis. *Int J Epidemiol* 2014;43:1865–73.
- [30] Puggina A, Broumas A, Ricciardi W, Boccia S. Cost-effectiveness of screening for lung cancer with low-dose computed tomography: a systematic literature review. *Eur J Public Health* 2015;5:1–8.
- [31] Gulati S, Mulshine JL. Lung cancer screening guidelines: common ground and differences. *Transl Lung Cancer Res* 2014;3:131–8.
- [32] Sharma D, Newman TG, Aronow WS. State of the art paper Lung cancer screening : history , current perspectives , and future directions. *Arch Med Sci* 2015;11:1033–43.
- [33] Broekhuizen H, Groothuis-Oudshoorn C, Vliegenthart R, Groen H, IJzerman M. Public preferences for lung cancer screening policies. *Value Heal* 2017.
- [34] Venerables WN, Ripley BD. Modern applied statistics with S. 4th ed. Springer; 2002.
- [35] Leisch F. FlexMix : A General Framework for Finite Mixture Models and Latent Class Regression in R. *J Stat Softw* 2004;11:1–18.
- [36] Prim RC. Short connection networks and some generalizations. *Bell Syst Tech J* 1957;36:1389–401.

- [37] Beasley J, Springer S. Algorithm AS 111: The Percentage Points of the Normal Distribution. *J R Stat Soc Ser C Applied Stat* 1977;26:118–21.
- [38] Lentz R. Computing noncentral beta probabilities. *Appl Stat* 1987;36:241–4.
- [39] Minka TP. Estimating a Dirichlet distribution. 2012.
- [40] Ripley BD. *Stochastic Models*. Stoch. Simul., Hoboken, NJ, USA: John Wiley & Sons, Inc; 1987, p. 98.
- [41] Cugliari J, Rolland A, Tran T. On the use of copulas to simulate multicriteria data. *DA2PL Work.*, 2014, p. 3–9.

Chapter 8

Assessing the value of lung cancer screening programs
under uncertainty in a heterogeneous population

Henk Broekhuizen

Catharina G.M. Groothuis-Oudshoorn

Rozemarijn Vliegenthart

Harry J.M. Groen

Maarten J. IJzerman

Abstract

Background: Lung cancer screening can reduce cancer-related mortality if screening program is well-designed and implemented. However, most implementation studies focus only on low-dose computed tomography (LDCT) and clinical attributes of screening. The aim of this study was to evaluate the perceived value of screening programs based on LDCT, breath analysis (BA), or blood biomarkers (BB) according to the perspective of the target population.

Methods: A multi-criteria decision analysis framework was adopted. The weights of seven attributes of screening were obtained from an earlier study that included a broad sample from the Netherlands. Performance data for the screening modalities was obtained from clinical trials and appended by expert opinion. Parameter uncertainty in the estimates of clinical performance was taken into account using a probabilistic approach, while heterogeneity in preferences was analyzed through subgroup analyses.

Results: The mean overall values were 0.58 (CI: 0.57 to 0.59), 0.57 (CI: 0.56 to 0.59), and 0.44 (CI: 0.43 to 0.45) for BB, BA, and LDCT, respectively. Seventy-seven per cent of respondents had a preference for BB or BA. For most subgroups, the overall values were similar to those of the entire sample. BA had the highest value for respondents who would have been eligible for earlier screening trials.

Discussion: The screening modalities BB and BA seem promising because they can be applied in a primary care setting. Although LDCT still seems preferable given its strong and positive evidence base, our results suggest that it is important to take non-clinical attributes into account to maximize attendance.

Introduction

Although the prognosis of cancer survival has greatly improved over the last decades, lung cancer mortality has not significantly reduced with the 5-year survival of only 18% [1]. Because the primary reason for the high mortality is that the disease has already metastasized once detected, it is found that early detection can substantially improve prognosis [2]. In the national lung screening trial (NLST) it was shown that annual screening with low-dose computed tomography (LDCT) can reduce cancer-related mortality by 20% compared to annual screening with chest x-ray [3]. Given these promising results, policy makers across the world are considering if lung cancer screening should be adopted as part of national screening programs [4]. However, there is still no consensus about actual screening guidelines or how these might be implemented [5]. A few studies have been performed investigating how lung cancer screening using LDCT should be implemented, yet these only focus on optimizing clinical parameters such as screening frequency or eligibility criteria [6,7]. Furthermore, although there is considerable experience with CT-based screening, recent advances in the fields of breath analysis (BA) and blood-biomarker (BB) screening modalities should be considered as well, because they might be more preferred by the general public due to convenience benefits [8–10]. Also, results from lung cancer screening trials suggest that there are some subgroups that are less likely to participate in screening after being invited than other subgroups [11,12]. This may be explained from previous research that shows that a decision to attend also depends on non-clinical attributes of screening such as comfort, duration and location [13] or on person-specific factors such as self-perceived risk of lung cancer [14]. Since low screening attendance rates may lead to lower cancer-related mortality reductions than predicted in trials, a lung cancer screening guideline should be optimized for attendance and thus appreciate both clinical diagnostic performance as well as other factors related to process and convenience of the modality [13].

Several survey methods such as multi-criteria decision analysis (MCDA) or discrete choice experiments (DCE) can be used to assess the relative value of hypothetical screening programs with different clinical and non-clinical attributes [15–17]. Examples of studies employing preference elicitation methods have been carried out in breast [18], colorectal [19–23], and prostate [24–26] cancer. The underlying assumption in these stated preference studies is that the screening program with the highest relative value is most preferred by the respondents, and therefore has the highest likelihood that the eligible population will attend. Although such preferences cannot directly predict screening attendance, they do provide insight into what attributes of screening programs influence

the intention to attend. In a previous study we identified seven attributes of screening programs for lung cancer that were most important for potential participants in the Netherlands using a stated preference survey [27]. The aim of the present study is to use MCDA to construct and evaluate the relative value for different screening programs that use either LDCT, BA, or BB. A second aim is to investigate whether the value of screening programs differs between preference subgroups or risk profile subgroups.

Methods

Multi-criteria decision analysis is a methodology stemming from operations research that can support a wide range of decisions. In general, a multi criteria decision analysis (MCDA) distinguishes four consecutive steps, i.e. structuring, weighting, scoring, and aggregating [28,29]. During the structuring step, the decision problem needs to be defined (who is the decision maker and what are the decision options) and important attributes of the decision alternatives are selected. The relative importance or attribute weight of each included attribute is then assessed using elicitation methods such as SWING weighting or MACBETH [28,30]. After this, the performance of each decision option on each attribute is measured and valued numerically in the scoring step. Then, the weights and scores are aggregated to obtain an overall value estimate for each decision option and the decision option with the highest value is considered the preferred choice. In the final step of the MCDA process the impact of various types of uncertainty in the scores, weights and overall values should be investigated and the results of the MCDA should then be communicated to stakeholders.

Structuring the decision

In this study, we investigated the decision whether or not to attend the first round of a screening program for lung cancer. We assumed the decision perspective of (potential) participants from a broadly determined eligible population in the Netherlands, which we assumed were all men and women between 40 and 80 years of age with no history of lung cancer. We assumed that someone is most likely to attend the screening program with the highest perceived value. Therefore, we aimed to assign a value to each alternative screening program using the respondents from a previously conducted stated preference study [27]. The program with the highest mean value across the respondent sample is denoted as the preferred program.

The screening programs in this study were based on three different modalities, namely low-dose computed tomography, breath analysis and blood biomarkers. In case of LDCT, a

malignancy is suspected if the diameter or volume (growth) of a nodule on a low-dose thorax image exceeds a predefined threshold [5]. We focused on LDCT with a threshold based on volume-doubling time instead of diameter-based malignancy threshold because a high false-positive rate was reported for the latter [3,31]. Regarding the second modality, breath analysis, there are multiple technologies [32]. We chose to focus on the electronic nose (“e-Nose”) technology [33]. With breath analysis one must exhale into a device similar to that used for lung function testing. The e-Nose detects (patterns of) volatile organic compounds in the exhaled air which is then compared to patterns found in lung cancer patients and healthy persons. If the chemical patterns in the collected breath sample match those of lung cancer patients, the participant is referred for further testing. For the last modality, a multitude of blood biomarker-based screening technologies exist [9,34,35]. For the present study, we chose to focus on circulating tumor cells in blood [36]. With these so-called liquid biopsies, blood is taken from the participant and a lab procedure is used to find malignant cells that have shed into the vascular system from the primary tumor.

We compared the three screening modalities on seven attributes of screening that were identified in a previous study [27]. The attributes that were identified together with interviewees and panel members were sensitivity, specificity, radiation burden, duration of screening process, waiting time until results are communicated, location of screening, and mode of screening. *Sensitivity* was defined as the probability of a positive (i.e. suspect) test result given that someone has lung cancer. *Specificity* was defined as the probability of a negative (i.e. non-suspect) test result given that someone does not have lung cancer. *Radiation burden* was defined as the radiation that a participant receives during a single screening on top of the normal yearly background radiation in the Netherlands (in mSV) [37]. The *duration of screening* was defined as the time the participant would spend at the facility where screening takes place (in minutes). *Time until results* was defined as the time until the participant receives the screening results, either through a consultation or with a letter sent to his/her home address (in days). *Mode of screening* was the modality of screening with ‘going through a scanner’, ‘exhaling into a device’, and ‘giving a blood sample’ as levels. The *location of screening* attribute has two levels: ‘the office of your general practitioner’ and ‘the hospital nearest to you’. For each attribute, a lower (‘worst’) and upper (‘best’) level was identified. The actual performance of the screening programs should fall in this range. Furthermore, it is assumed that respondents are indifferent to performance changes outside of this range, e.g. performing worse than the lower level confers the same value as performance exactly at the lower level. For the sensitivity, specificity, and radiation burden attributes the lower and upper levels were identified based on literature and the clinical expertise of authors RV and HG.

	Attribute name	Worst level	Best level
Continuous attributes	Sensitivity	70%	100%
	Specificity	70%	100%
	Radiation burden	Background + 1.5 mSv	Background
	Duration of screening procedure	45 minutes	15 minutes
	Time until results	14 days	1 day
Categorical attributes	Scan type	levels: 'lie in scanner', 'sustained breath into device', 'give blood'	
	Location	levels: 'nearest hospital', 'at your GP'	

Table 1: Included attributes and levels. Sensitivity is the probability of a positive/suspect test results given that the person has lung cancer. Specificity is the probability of a negative test results given that the person does not have lung cancer. The worst and best levels for the continuous attributes were defined based on interviews with clinicians. For the categorical attributes the preference order was elicited from respondents in separate questions.

In the preference survey the percentages used for sensitivity and specificity were rounded to the nearest single decimal percentage (i.e. 10%, ..., 100%) to improve clarity for respondents. The worst and best levels for the duration of screening and waiting time until results attributes was established in the panel session. Finally, the preferential order of mode and modality of screening cannot be assumed *a priori* (e.g. it will differ between persons whether the GP location is preferred over the hospital location). Respondents in the preference survey were therefore asked to rank the modes of screening from some least burdensome to most burdensome and they were asked to indicate their most preferred location. An overview of the attributes and their levels is given in Table 1.

Weighting attributes using a preference survey

The aim of the weighting step is to identify the relative importance of attributes with attribute weights. These attribute weights were elicited in the preference study from the respondents using a swing weighting task [27,38]. From this we obtained a set of seven attribute weights for each respondent. These weights indicate the relative importance of improvements from the worst performance level to the best performance level between attributes [39]. One thousand and thirty-four respondents from the Dutch general population completed the preference survey. Fifty one percent of the respondents were female, the mean age was 58 (SD=11), and the respondents were distributed evenly between low, medium and higher levels of education. For more details regarding the

<i>Attribute name</i>	Mean weight (SD)	Bisection point percentiles		
		2.5%	50%	97.5%
<i>Sensitivity</i>	0.16 (0.14)	63.5%	90.0%	100.0%
<i>Specificity</i>	0.13 (0.12)	64.3%	90.4%	100.0%
<i>Radiation burden</i>	0.13 (0.12)	0.16 mSv	1.10 mSv	1.5 mSv
<i>Duration of screening procedure</i>	0.10 (0.09)	42.0 min	26.4 min	15.0 min
<i>Waiting time until results</i>	0.13 (0.12)	12.8 days	5.2 days	1 days
<i>Mode of screening</i>	0.17 (0.14)	21.0%	72.5%	100.0%
<i>Location of screening</i>	0.18 (0.17)	-	-	-

Table 2: Preference data. For mode of screening, the bisection point indicates the respondent's indifference between an improvement from the worst level to the bisection point and the improvement from the bisection point to the best level. No bisection point was elicited for the location of screening because it had only two (categorical) levels. The lower and upper levels per attribute can be found in Table 1. More details can be found in [27].

attribute-selection process, the preference elicitation process and the characteristics of the population sample, the reader is referred to [27].

Measuring the performance data of screening programs

The aim of the scoring step is to find out how well screening programs perform on the included attributes. To be able to do this, we collected clinical evidence on the screening programs' performances. Where this was not available, we made assumptions based on the clinical experience of the authors RV and HG. For the sensitivity and specificity of LDCT published pooled estimates from the first round of the NELSON trial were used [40]. For the sensitivity and specificity of BA we included the trials from a recent systematic review [32]. For the estimation of the predictive power of BA in terms of the sensitivity and specificity a prediction model is constructed based on the chemical patterns found in known cancer patients and known healthy persons. Prediction models constructed in this way are susceptible to 'overfitting', an undesirable effect where the model memorizes the persons in the case control study instead of identifying generalizable patterns and this leads poor predictive power for new cases outside of the original case control study. We excluded trials from the review that did not correct for overfitting as their performance estimates (mostly very close to 100%) were unlikely to generalize well [32]. We subsequently pooled the sensitivity and specificity estimates from the included eight clinical trials with a bivariate copula random effects model where the choice of copula function was made with maximum likelihood estimation [41]. Performance data for sensitivity and specificity of BB

was also obtained from trials cited in a systematic review [36]. As for BA, the estimates for BB were pooled using a bivariate copula random effects model [41]. The radiation burden for LDCT was set to 1.5 mSv and the radiation burden for BA and BB was set to 0 mSv. For duration of screening we assumed that the participant would need between 15 to 45 minutes for LDCT, between 20 to 30 minutes for BA, and between 10 to 25 minutes for BB. LDCT images have to be evaluated by a radiologist and we assumed it would take one week for the participant to receive the result. Samples for BB are mostly analyzed in separate lab facilities and therefore the time until results for BB was set to two days. We assumed the BA results would be available the next day because the BA review authors noted that “*time between breath collection and analysis was usually a couple of hours*” [32]. We assumed that LDCT screening is performed at the local hospital. Screening based on BA or BB is assumed to be done at a primary care facility.

Finally, the three screening technologies directly corresponded to the levels for the mode of screening attribute: for LDCT one ‘goes through a scanner’, for BA one ‘has to exhale into a device’, and for BB one has to ‘give blood’. An overview of all performance inputs can be found in Table 3.

<i>Attribute name</i>	LDCT	Breath analysis	Blood biomarkers
<i>Sens*</i> (%)	92.5 (85.3 to 97.5)	80.0 (52.8% to 96.7)	74.8 (66.7 to 85.2)
<i>Spec†</i> (%)	98.8 (98.0 to 98.6)	79.8 (66.7% to 89.9)	87.9 (77.1 to 95.7)
<i>Rad</i> (mSv)	1.5	0	0
<i>Dur</i> (minutes)	30 (range: 15 to 45)	25 (range: 20 to 30)	17.5 (range 10 to 25)
<i>Wait</i> (days)	7	1	2
<i>Mode</i>	Go through scanner	Exhale into device	Give blood
<i>Location</i>	Hospital	GP’s office	GP’s office

Table 3: Used performance data of the three included screening policies with 95% confidence intervals (for sens and spec) or ranges (for dur). *pooled with Frank copula with Kendall’s $\tau = 0.9$, †pooled with normal copula with Kendall’s $\tau = 1.0$. sens=sensitivity, spec=specificity, rad=radiation burden, dur=duration of screening, time=waiting time until results.

Aggregating into overall values

In the aggregation step, a value score is calculated for each screening program. Since performances are measured on different natural scales (percentages, days, *etc.*) we first need to transform these metrics to a partial value between zero (no value) and one (maximum value). Finally, the value estimate for each included screening policy is obtained by weighting the partial values with the attribute weights.

A partial value function is a mathematical function that maps attribute-specific performances (percentages, days, *etc.*) onto a partial value scale between zero and one. Although partial value functions are commonly assumed to be linear, we considered this to be unrealistic in our case as there are two time-related attributes in our attribute set and people's preferences for time are often non-linear [42]. Moreover, we considered it unlikely that the difference in preference between the first and second ranked screening mode was equal to the difference in preference between the second and third ranked screening mode.

To estimate a non-linear partial value function for the continuous attributes we elicited from each respondent a bisection point on the performance scale of each attribute (except for location). We excluded 48 respondents who did not complete the bisection point questions. The bisection point indicates the respondent's indifference between an improvement from the worst level to the bisection point and the improvement from the bisection point to the best level. Denoting the indifference point with x_{kq}^* and the worst and best levels with x_k^- and x_k^+ respectively, we obtained the two-piece partial value function v_{kq} of continuous attribute k ($k = 1, \dots, 5$) for respondent q ($q = 1, \dots, n$):

$$v_{kq}(x_{ki}; x_{kq}^*, x_k^-, x_k^+) = \begin{cases} 0, & \text{if } x_{ki} \leq x_k^- \\ \frac{1}{2} \left(\frac{x_{ki} - x_k^-}{x_{kq}^* - x_k^-} \right), & \text{if } x_k^- < x_{ki} < x_{kq}^* \\ \frac{1}{2} \left(1 + \frac{x_{ki} - x_{kq}^*}{x_k^+ - x_{kq}^*} \right), & \text{if } x_{kq}^* \leq x_{ki} < x_k^+ \\ 1, & \text{if } x_{ki} \geq x_k^+ \end{cases} \quad (1)$$

Here, x_{ki} is the performance of screening program i on attribute k . The sixth attribute, mode of screening, is categorical. Therefore, we defined the partial value $v_{6,q}$ for the third ranked mode for a respondent q as having zero value, the second rank mode as having $x_{6,q}^*$ value (the bisection point) and the first ranked mode as having a value of one. The seventh attribute (location of screening) is also categorical with two levels so the partial value $v_{7,q}$ for the second ranked location was zero and the partial value for the first ranked location was one.

To aggregate the respondent's partial values for the different attributes into an overall value for each included screening program (denoted V_{iq}), the partial values were scaled using the attribute weights and then summed:

$$V_{iq} = \sum_{k=1}^7 w_{kq} v_{kq}(x_{ki}; x_{kq}^*) \quad (2)$$

Then, an estimate of the mean overall value for a screening program in the general population, denoted V_i , was obtained by averaging the individual overall values over all respondents:

$$V_i = \frac{\sum_{q=1}^n V_{iq}}{n} \quad (3)$$

By sorting the screening programs per respondent from highest value to lowest value a rank order of screening programs per respondent r_{iq} was obtained. Furthermore, we denoted the rank of program i with R_i . The proportion of the population for who this rank equals $y = 1, 2, \text{ or } 3$ was then estimated with the proportion of respondents for which program i had rank y (Formula's 4 and 5):

$$\frac{\sum_{q=1}^n 1_y(r_{iq})}{n} \quad (4)$$

where

$$1_y(r_{iq}) = \begin{cases} 1, & \text{if } r_{iq} = y \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Uncertainty analyses

We investigated the impact of parameter uncertainty in clinical evidence and heterogeneity in attribute weights on the mean overall values of the screening programs [43,44]. Parameter uncertainty was investigated using a probabilistic approach. The uncertain parameter estimators in our model were those for the programs' performance on the sensitivity, specificity, and duration of screening attributes. For LDCT, we modeled the uncertainty in the parameter estimates for sensitivity and specificity by two independent beta distributions according to best practices for modeling proportions [45]. We also assumed that the average screening duration for LDCT was uniformly distributed between 15 and 45 minutes. For both BA and BB the uncertainty in the estimators for sensitivity and specificity was modelled with the bivariate copula distributions that were fitted for the pooling of study-level estimates [41]. Furthermore, we assumed that the average screening duration was uniformly distributed between 20 to 30 minutes for BA and between 10 to 25

minutes for BB. Since the overall value of each screening program is a combination of weights and performances, it is hard to analytically find the distribution of V_i given the parameter uncertainty in sensitivity, specificity, and duration. We therefore estimated this value distribution using Monte Carlo simulations [46]. This meant that we drew one sample from the aforementioned probability distributions of the performance estimators for each respondent and then calculated the partial values of each program for each respondent with Formula's 1a – 1c and the overall values using Formula 2.

We also calculated the 95% confidence intervals of the partial and overall values across the preference sample by taking the empirical 95% confidence interval from the Monte Carlo simulations. An overview of the Monte Carlo process can be found in Figure 1.

Figure 1: Modeling flowchart.

To investigate if heterogeneity in preferences would influence which of the screening programs had the highest value, we performed the following subgroup analyses. Each scenario consisted of a subgroup from the preference study. First, with cluster analysis the respondents were divided into five subgroups of persons with similar preferences [27]. Each of these preference subgroups is named after the attribute(s) respondents in the subgroup considered most important: ‘radiation-sensitivity’ (n=236), ‘waiting time until results’ (n=157), ‘location-mode’ (n=299), ‘mode-sensitivity’ (n=184), and ‘sensitivity-specificity’ (n=158). Apart from these preference subgroups, we also investigated subgroups in the preference study that can be defined in terms of their eligibility for screening according to three recent screening trials. Each of these trials (NLST, NELSON, and UKLS) had different eligibility criteria based on a person’s risk profile. The NLST study included men and women aged between 55 and 74 who had more than 30 pack-years of smoking [3]. Furthermore, former smokers must have quit within the previous 15 years. A pack-year is defined as twenty cigarettes smoked every day for one year. In the preference study, it was not asked how many cigarettes respondents smoked per day but instead it was asked how many years respondents had smoked and if they had quit, how long ago that was. Since the average number of cigarettes smoked in the Netherlands by smokers in the eligible population is 11 [47], we multiplied the total number of years smoked by $11/20=0.55$ to estimate the respondents’ pack years. Here we also assume that the daily number of cigarettes smoked by someone is constant. The NELSON study included men and women aged between 50 and 78 that smoked either more than 15 cigarettes per day during more than 25 years or more than 10 cigarettes/day during more than 30 years [48]. Furthermore, to be eligible for the NELSON study one needs to be a current smoker or a former smoker who quit within the previous 10 years. Finally, the UKLS study included men and women aged between 50 and 75 with 5-year lung cancer risk of more than 5% as calculated with the Liverpool Lung Project model [49,50]. Note that the subgroups based on the screening trials were not mutually exclusive: i.e. persons with a very high risk of lung cancer are likely to be eligible for multiple trials.

For each of the eight subgroups we re-ran the model, including parameter uncertainty in estimates for sensitivity, specificity, and duration of screening. In this way we obtained mean value estimates and associated confidence intervals of the screening programs according to each subgroup. By comparing these mean values we assessed which screening program would be most valuable for different subgroups.

Results

The partial values of the screening programs for sensitivity and specificity reflected the differences in the clinical evidence: LDCT had the highest partial value for both sensitivity and specificity, BA had a higher partial value for sensitivity than BB, and BB had a higher partial value for specificity than BA. LDCT had the lowest partial value for the duration of screening and waiting time until results attributes, BB had the highest partial value for the duration of screening attribute, and BA had the highest partial value for the waiting time until results attribute. LDCT and BB both had similar partial values for mode of screening (around 0.65) while BA had a clearly lower partial value on that attribute (0.37). Because BA and BB were both assumed to take place at the GP’s office and both had no radiation burden, the partial values for these attribute was identical.

	LDCT	Breath analysis	Blood biomarkers
<i>Sens</i>	0.63 (0.62 to 0.64)	0.39 (0.37 to 0.41)	0.33 (0.32 to 0.34)
<i>Spec</i>	0.81 (0.80 to 0.82)	0.37 (0.36 to 0.38)	0.52 (0.51 to 0.53)
<i>Rad</i>	0.00 (0.00 to 0.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
<i>Dur</i>	0.17 (0.16 to 0.18)	0.23 (0.23 to 0.23)	0.53 (0.51 to 0.55)
<i>Wait</i>	0.26 (0.26 to 0.26)	1.00 (1.00 to 1.00)	0.45 (0.45 to 0.45)
<i>Mode</i>	0.65 (0.63 to 0.67)	0.37 (0.34 to 0.40)	0.68 (0.65 to 0.71)
<i>Location</i>	0.40 (0.37 to 0.43)	0.60 (0.57 to 0.63)	0.60 (0.57 to 0.63)

Table 4: mean partial values (as calculated for the base case with Formula 1). 95% confidence interval were obtained with Monte Carlo simulations. sens=sensitivity, spec=specificity, rad=radiation burden, dur=duration of screening, time=waiting time until results.

BB had a mean overall value of 0.58 (CI: 0.57 to 0.59) and BA had a mean overall value of 0.57 (CI: 0.56 to 0.59). Although the mean overall value for BB was slightly higher than that of BA, the absolute difference is small (0.01) and the confidence intervals due to parameter uncertainty overlap. LDCT, however, clearly had the lowest mean overall value 0.44 and its confidence interval that runs from 0.43 to 0.45 does not overlap with either that of BB or BA. Although the distribution of the overall value distributions are skewed to the left the median overall values coincide with the mean. In contrast with the confidence intervals around the mean overall values estimates, which were small, the individual value distributions showed considerable overlap. This is reflected in the rank proportions: BB had the highest value for 43.9% of respondents, while this was 32.7% for BA and 23.4% for LDCT. For 56.9% of respondents LDCT had the last rank. An overview of the overall values and the rank probabilities is presented in Table 5. For most preference subgroups, the

overall values were similar to those of the entire sample (Table 7 and Figure 2). For the respondents in the ‘waiting time until results’ preference subgroup however, breath analysis had the highest value and its confidence interval did not overlap that of the other screening programs. In the ‘sensitivity-specificity’ preference subgroup, LDCT had the highest mean value although its confidence interval overlapped with that of the other programs. The preference for BB according to respondents in the ‘radiation-sensitivity’

Parameter uncertainty		Stochastic uncertainty					
		Value percentiles			Per cent of respondents for who modality was ranked ...		
	Mean overall value (CI)	5%	50%	95%	First	Second	Third
<i>LDCT</i>	0.44 (0.43 to 0.45)	0.16	0.44	0.74	23.4%	19.7%	56.9%
<i>BA</i>	0.57 (0.56 to 0.59)	0.26	0.58	0.87	32.7%	45.0%	22.3%
<i>BB</i>	0.58 (0.57 to 0.59)	0.26	0.60	0.87	43.9%	35.3%	20.8%

Table 5: Overview of values and rank proportions. The confidence interval for the mean overall values reflects uncertainty on the sample level due to uncertainty in parameter estimates for clinical performance. The results under ‘stochastic uncertainty’ reflect differences on the respondent level.

preference subgroup was stronger than that of the entire sample in that the confidence interval for BB no longer overlapped with that of BA. From the sample 10.3 to 18.1 per cent was eligible for one of the trials (see Table 6). The attribute weights as assigned by respondents in these particular subgroups were similar to those of the whole sample. However, they seemed to consider the burden of exhaling into a device more severe and giving blood less severe than the sample as a whole [27]. Also, whereas in the entire sample 60% prefers a GP’s office to a hospital, this is only 53% for those eligible for the NLST trial. In the subgroup of respondents who are eligible for the UKLS trial, 48% of respondents prefer the GP. With regard to the overall values for these risk subgroups, we see that the overall values for the NLST and NELSON subgroups are similar to those of the entire sample, i.e. BA and BB have similar values and LDCT has a significantly lower values. For the respondents who were eligible for UKLS, BA had the highest value although the confidence intervals of all three programs overlapped.

Discussion

In this study, we have assessed the relative value of three lung cancer screening policies using a probabilistic MCDA model. BB had the highest mean overall value but the difference

between BB and BA is so small that it is unlikely to have policy relevance. We do see that LDCT had the lowest mean overall value and this is stable given the parameter uncertainty in estimates for clinical performance. Furthermore, LDCT had the lowest mean overall value in most subgroups we investigated. It is important to consider that the current study was based on a public sample, from which only 10 to 18 per cent is currently eligible for screening under guidelines used by recent lung cancer screening trials. If only the preferences of those eligible for screening according to the criteria used in the NELSON or NLST trial are taken into account, it can be seen that the overall values for the screening programs are very similar to those found in the entire sample. For the respondents who are eligible for screening according to the UKLS trial criteria, BA has the highest mean value but the confidence intervals around the mean overall values for all three screening programs overlap. Although our analyses show that the mean values for BA and BB are higher than for LDCT, there is still considerable variation in the value of screening programs for individual respondents as can be seen in the ranking proportions.

Figure 2: Overall mean values for LDCT (blue), breath analysis (green), and blood biomarkers (red) with 95% confidence interval. ‘Base’ denotes the total sample, PSG x denotes the x th preference subgroup (see Table 6 for an overview).

It seems that the majority of respondents focus on non-clinical attributes, while only 15% of respondents (namely those in the ‘sensitivity-specificity’ subgroup) seemed to focus on diagnostic accuracy. Unsurprisingly, LDCT had the highest mean value for this subgroup. For the other subgroups and the sample as a whole, LDCT had the lowest mean overall value. This implies that non-clinical attributes of screening play an important role in people’s preferences for screening programs, explaining why LDCT has such a low mean value. However, the high mean values for BA and BB found in our study may

			Mean attribute weight (SD)							
			n (%)	Se	Sp	Ra	Du	Wa	Mo	Lo
Preference subgroup	PSG1	227 (23)	0.21 (0.12)	0.20 (0.11)	0.25 (0.17)	0.09 (0.07)	0.07 (0.05)	0.12 (0.10)	0.07 (0.05)	
	PSG2	174 (18)	0.15 (0.11)	0.10 (0.06)	0.09 (0.06)	0.11 (0.11)	0.11 (0.06)	0.31 (0.20)	0.12 (0.07)	
	PSG3	284 (29)	0.08 (0.06)	0.06 (0.04)	0.10 (0.06)	0.10 (0.08)	0.11 (0.07)	0.18 (0.12)	0.36 (0.18)	
	PSG4	151 (15)	0.09 (0.06)	0.07 (0.04)	0.14 (0.09)	0.13 (0.12)	0.29 (0.18)	0.14 (0.08)	0.14 (0.07)	
	PSG5	150 (15)	0.30 (0.19)	0.26 (0.17)	0.07 (0.05)	0.08 (0.07)	0.11 (0.07)	0.09 (0.05)	0.09 (0.06)	
Population according to	NLST	188 (18)	0.15 (0.14)	0.12 (0.10)	0.12 (0.10)	0.10 (0.09)	0.14 (0.12)	0.18 (0.16)	0.18 (0.16)	
	NELSON	183 (18)	0.16 (0.15)	0.13 (0.12)	0.11 (0.10)	0.10 (0.09)	0.14 (0.12)	0.17 (0.16)	0.18 (0.15)	
	UKLS	106 (10)	0.15 (0.14)	0.12 (0.11)	0.10 (0.08)	0.10 (0.08)	0.14 (0.13)	0.20 (0.18)	0.18 (0.16)	

Table 6a: Overview of the scenarios. The preference subgroups names are based on the two most important attributes in that cluster [27]. The mean weights with standard deviations are presented per subgroup as well. Se=sensitivity, Sp=specificity, Ra=radiation burden, Du=duration of screening, Wa=waiting time until results, Mo=mode of screening, Lo=location of screening, FM=first ranked mode, FL=first ranked location, psgx=preference subgroup number x; here 1: 'Radiation-sensitivity', 2: 'Mode-sensitivity', 3: 'Location-mode', 4: 'Waiting time until results', 5: 'Sensitivity-specificity'.

		First ranked mode (% of respondents in subgroup)			First ranked location (% of respondents in subgroup)	
		LDCT	BA	BB	GP	
Preference subgroup	PSG1	16	59	25	57	
	PSG2	32	34	34	61	
	PSG3	23	39	38	65	
	PSG4	28	46	26	56	
	PSG5	21	50	29	55	
Population according to	NLST	25	38	37	53	
	NELSON	20	43	37	55	
	UKLS	28	30	42	48	

Table 6b: Overview of the scenarios. The preference subgroups names are based on the two most important attributes in that cluster [27]. psgx=preference subgroup number x; here 1: 'Radiation-sensitivity', 2: 'Mode-sensitivity', 3: 'Location-mode', 4: 'Waiting time until results', 5: 'Sensitivity-specificity'.

be biased by the low quality level of evidence. No prospective clinical trials have been done comparing BA or BB to LDCT, and the diagnostic accuracy estimates we use are for clinical tumors, not pre-clinical tumors that would be more relevant in a screening setting. In contrast, multiple large prospective screening trials have been performed for LDCT, which provides confidence and a solid and positive argument for its use in screening. It would therefore be better to consider our results in light of the question of how a screening program based on LDCT can be further improved and implemented.

Our results can support policy makers define strategies that can help the screening program reach as many people as possible. Two practical concerns that reduce the value of LDCT are waiting time until results and radiation burden. An improvement strategy may thus be to try and reduce waiting time until results are communicated to people, for example by evaluating CT images using machine learning algorithms [51]. Secondly, the fact that the respondents in our sample seemed quite worried about the radiation burden posed by LDCT (mean weight of 0.13) is interesting because the actual incremental risk due to the radiation induced by LDCT is likely to be small. A strategy for addressing these concerns may be targeted information campaigns that inform and shift people's views. If machine learning could reduce the waiting time until results to one day and if the information campaign would reduce the radiation concerns, our model shows that the overall value for LDCT would be close or equal to BA and BB. This would be beneficial because it improves people's satisfaction with the screening program, subsequently leading to more people attending first or follow-up screening rounds. This increased attendance would increase the cost-effectiveness of the screening program as a whole.

	LDCT	Breath analysis	Blood biomarkers
<i>'Radiation-sensitivity'</i>	0.44 (0.42 to 0.46)	0.57 (0.55 to 0.59)	0.63 (0.61 to 0.65)
<i>'Mode-sensitivity'</i>	0.47 (0.45 to 0.49)	0.56 (0.53 to 0.59)	0.55 (0.52 to 0.58)
<i>'Location-mode'</i>	0.42 (0.39 to 0.44)	0.58 (0.56 to 0.60)	0.61 (0.58 to 0.64)
<i>'Waiting time until results'</i>	0.36 (0.35 to 0.38)	0.66 (0.64 to 0.68)	0.58 (0.56 to 0.60)
<i>'Sensitivity-specificity'</i>	0.54 (0.52 to 0.56)	0.50 (0.47 to 0.52)	0.51 (0.49 to 0.53)
<i>NLST inclusion criteria</i>	0.47 (0.44 to 0.50)	0.56 (0.53 to 0.59)	0.53 (0.51 to 0.56)
<i>NELSON inclusion criteria</i>	0.47 (0.45 to 0.50)	0.55 (0.52 to 0.57)	0.55 (0.52 to 0.58)
<i>UKLS inclusion criteria</i>	0.51 (0.47 to 0.54)	0.54 (0.50 to 0.58)	0.51 (0.47 to 0.54)

Table 7: Overall values (with 95% confidence intervals) for the three different screening programs according to each of the included subgroups.

The main strength of our study is that we had a large representative sample from the Dutch population where we could study the value of screening programs for subgroups based on preferences or based on eligibility for trials. Second, by eliciting bisection points we improved upon the commonly used but unrealistic assumption of linear partial value functions in MCDA for healthcare. Third, we provide an evidence-based assessment of the perceived value of lung cancer screening policies by explicitly including clinical evidence and results from a stated preference study. As stated before, the main limitation of our study is that the quality of evidence for BA and BB was low. This precludes us from making strong statements about the value of these modalities in screening practice. Moreover, because we did not know the pack-years of the respondents but only years of smoking, we had to make assumptions regarding people's smoking habits based on publicly available data. Finally, we have taken the perspective of the potential participant who is deciding which screening policy to attend for the first time. We have not taken into account the long term consequences of screening or perceived values relevant for subsequent screening rounds.

In conclusion, we have evaluated the value of three different screening policies by combining data from a large public preference sample and clinical evidence. Although the evidence base for lung cancer screening based on breath analysis or blood biomarkers is still weak, these screening modalities seem promising because of their practical advantages like application in a primary care setting. To improve the value of LDCT-based screening for the eligible population it is important to take non-clinical attributes of screening into account.

References

- [1] Optican RJ, Chiles C. Implementing lung cancer screening in the real world: opportunity, challenges and solutions. *Transl Lung Cancer Res* 2015;4:353–64.
- [2] Kauczor H-U, Bonomo L, Gaga M, Nackaerts K, Peled N, Prokop M, et al. ESR/ERS white paper on lung cancer screening. *Eur Respir J* 2015;46:28–39.
- [3] The national lung screening trial research team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
- [4] Molina JR, Yang P, Cassivi SD, Schild SE, Adjei A a. Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clin Proc* 2008;83:584–94.
- [5] Gulati S, Mulshine JL. Lung cancer screening guidelines: common ground and differences. *Transl Lung Cancer Res* 2014;3:131–8.

- [6] McMahon PM, Meza R, Plevritis SK, Black WC, Tammemagi CM, Erdogan A, et al. Comparing benefits from many possible computed tomography lung cancer screening programs: Extrapolating from the National Lung Screening Trial using comparative modeling. *PLoS One* 2014;9.
- [7] de Koning HJ, Meza R, Plevritis SK, Ten Haaf K, Munshi VN, Jeon J, et al. Benefits and Harms of Computed Tomography Lung Cancer Screening Strategies: A Comparative Modeling Study for the U.S. Preventive Services Task Force. *Ann Intern Med* 2013.
- [8] Adiguzel Y, Kulah H. Biosensors and Bioelectronics Breath sensors for lung cancer diagnosis. *Biosens Bioelectron* 2015;65:121–38.
- [9] Xiang D, Zhang B, Doll D, Shen K, Kloecker G, Freter C. Lung cancer screening: from imaging to biomarker. *Biomark Res* 2013;1:4.
- [10] Chan HP, Lewis C, Thomas PS. Exhaled breath analysis: Novel approach for early detection of lung cancer. *Lung Cancer* 2009;63:164–8.
- [11] Ford ME, Havstad SL, Flickinger L, Johnson CC. Examining the effects of false positive lung cancer screening results on subsequent lung cancer screening adherence. *Cancer Epidemiol Biomarkers Prev* 2003;12:28–33.
- [12] Montes U, Seijo LM, Campo a., Alcaide a. B, Bastarrika G, Zulueta JJ. Factors determining early adherence to a lung cancer screening protocol. *Eur Respir J* 2007;30:532–7.
- [13] Mansfield C, Tangka FKL, Ekwueme DU, Smith JL, Guy GP, Li C, et al. Stated Preference for Cancer Screening: A Systematic Review of the Literature, 1990-2013. *Prev Chronic Dis* 2016;13:E27.
- [14] Wildstein KA, Faustini Y, Yip R, Henschke CI, Ostroff JS. Longitudinal predictors of adherence to annual follow-up in a lung cancer screening programme. *J Med Screen* 2011;18:154–9.
- [15] Weernink MGM, Janus SIM, van Til J a., Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceut Med* 2014.
- [16] Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl Health Econ Health Policy* 2013;11:319–29.
- [17] Tony M, Wagner M, Khoury H, Rindress D, Papastavros T, Oh P, et al. Bridging health technology assessment (HTA) with multicriteria decision analyses (MCDA): field testing of the EVIDEM framework for coverage decisions by a public payer in Canada. *BMC Health Serv Res* 2011;11:329.
- [18] Phillips KA, Van Bebber S, Marshall D, Walsh J, Thabane L. A review of studies examining stated preferences for cancer screening. *Prev Chronic Dis* 2006;3:A75.
- [19] Wortley S, Wong G, Kieu A, Howard K. Assessing stated preferences for colorectal cancer screening: a critical systematic review of discrete choice experiments. *Patient* 2014;7:271–82.
- [20] Hummel JM, Steuten LGM, Groothuis-Oudshoorn CJM, Mulder N, IJzerman MJ. Preferences for colorectal cancer screening techniques and intention to attend: A multi-criteria decision analysis. *Appl Health Econ Health Policy* 2013;11:499–507.
- [21] Groothuis-Oudshoorn CGM, Fermont J, van Til JA, IJzerman MMJ, et al. Public stated preferences and predicted uptake for genome-based colorectal cancer screening. *BMC Med Inform Decis Mak* 2014;14:18.
- [22] Ghanouni A, Smith S, Halligan S, Plumb A, Boone D, Yao G, et al. Public preferences for colorectal cancer screening tests: a review of conjoint analysis studies. *Expert Rev Med Devices* 2013;10:489–99.
- [23] Marshall D, McGregor E, Currie G. Measuring Preferences for Colorectal Cancer Screening. *Patient Patient-Centered Outcomes Res* 2010;3:79–89.

- [24] de Bekker-Grob EW, Rose JM, Donkers B, Essink-Bot M-L, Bangma CH, Steyerberg EW. Men's preferences for prostate cancer screening: a discrete choice experiment. *Br J Cancer* 2013;108:533–41.
- [25] Kalkeld G, Patel M, Mann G, Pignone M. Men's preferences and trade-offs for prostate cancer screening: a discrete choice experiment. *Heal Expect* 2014.
- [26] Kaltoft M, Turner R, Nielsen J, Cunich M, Salkeld G, Dowie J. Addressing preference heterogeneity in public policy by combining Cluster Analysis and Multi-Criteria Decision Analysis. *Health Econ Rev* 2014;5:1–11.
- [27] Broekhuizen H, Groothuis-Oudshoorn C, Vliegenthart R, Groen H, IJzerman M. Public preferences for lung cancer screening policies. *Value Health* 2017.
- [28] Dodgson J, Spackman M, Pearman A, Phillips L. Multi-criteria analysis : A manual. London: 2009.
- [29] Marsh K, IJzerman M, Thokala P, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making —Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Heal* 2016;19:1–13.
- [30] Bana E Costa C a., De Corte J-M, Vansnick J-C. Macbeth. *Int J Inf Technol Decis Mak* 2012;11:359–87.
- [31] Xu DM, Gietema H, de Koning H, Vernhout R, Nackaerts K, Prokop M, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer* 2006;54:177–84.
- [32] Krilaviciute A, Heiss J, Leja M, Kupcinskas J, H H, Brenner H. Detection of cancer through exhaled breath: a systematic review. *Oncotarget* 2015;6:38643–57.
- [33] Adiguzel Y, Kulah H. Breath sensors for lung cancer diagnosis. *Biosens Bioelectron* 2015;65:121–38.
- [34] Sozzi G, Boeri M. Potential biomarkers for lung cancer screening. *Transl Lung Cancer Res* 2014;3:139–48.
- [35] Zhang Z, Ramnath N, Nagrath S. Current Status of CTCs as Liquid Biopsy in Lung Cancer and Future Directions. *Front Oncol* 2015;5:1–11.
- [36] Hamilton G, Rath B. Detection of circulating tumor cells in non-small cell lung cancer. *J Thorac Dis* 2016;8:1024–8.
- [37] Eleveld H. Ionising radiation exposure in the Netherlands. Bilthoven: 2003.
- [38] Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making-An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Heal* 2015;19:1–13.
- [39] Belton V, Stewart TJ. Multiple criteria decision analysis : An integrated approach. 2nd ed. Dordrecht: Kluwer Academic Publishers; 2002.
- [40] Horeweg N, Scholten ET, de Jong P a., van der Aalst CM, Weenink C, Lammers JWJ, et al. Detection of lung cancer through low-dose CT screening (NELSON): A prespecified analysis of screening test performance and interval cancers. *Lancet Oncol* 2014;15:1341–50.
- [41] Nikoloulopoulos AK. A mixed effect model for bivariate meta-analysis of diagnostic test accuracy studies using a copula representation of the random effects distribution. *Stat Med* 2015;34:3842–65.
- [42] Andreoni J, Kuhn M, Sprenger C. Measuring time preferences: A comparison of experimental methods. *J Econ Behav Organ* 2015;116:451–64.
- [43] Broekhuizen H, Groothuis-Oudshoorn CGM, van Til JA, Hummel JM, IJzerman MJ. A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics* 2015;33:445–55.

- [44] Groothuis-Oudshoorn C, Broekhuizen H, van Til J. Dealing with Uncertainty in the Analysis and Reporting of MCDA. In: Marsh K, Goetghebeur M, Thokala P, Baltussen R, editors. *Multi-Criteria Decis. Anal. to Support Healthc. Decis.* 1st ed., Springer; 2017, p. 65–83.
- [45] Briggs AH, Weinstein MC, Fenwick E a L, Karnon J, Sculpher MJ, Paltiel a D. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Heal* 2012;15:835–42.
- [46] Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber AB, Jansen JP, IJzerman MJ. Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis. *BMC Med Inform Decis Mak* 2015;15:1–10.
- [47] Health survey data [Data van gezondheidsenquête]. Stat Netherlands 2016.
- [48] Oudkerk M, Heuvelmans MA. Screening for lung cancer by imaging: the Nelson study. *JBR-BTR* 2013;96:163–6.
- [49] Baldwin DR, Duffy SW, Wald NJ, Page R, Hansell DM, Field JK. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. *Thorax* 2011;66:308–13.
- [50] Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;98:270–6.
- [51] Demir Ö, Yılmaz Çamurcu A. Computer-aided detection of lung nodules using outer surface features. *Biomed Mater Eng* 2015;26:S1213–22.

Chapter 9

General Discussion

Main findings

The aim of this thesis was to investigate how the various types of uncertainty influence the comparison of medical interventions in a value-framework incorporating stakeholder preferences using multi-criteria decision analysis (MCDA). Because of the focus on uncertainty, the thesis started with a review of methods to incorporate uncertainty (second chapter) in MCDA models. In that chapter, we identified five approaches for dealing with uncertainty in MCDA and concluded that while deterministic sensitivity analysis is preferred for reasons of simplicity, the probabilistic approach was the better method when there are multiple (possibly correlated) sources of uncertainty in either clinical evidence or stakeholder preferences. We then applied the probabilistic approach to two cases, namely antidepressants (chapter 3) and antiretroviral treatments for HIV (chapter 4). In both chapters we introduced an MCDA model that can be used for weighting clinical evidence using patient preferences and we illustrated that the investigated decisions were sensitive to both uncertainty in preferences and clinical performances.

The role of uncertainty in stakeholder preferences was then further explored in the second half of the thesis specifically in relation to the implementation of lung cancer screening. In chapter 5 we explore this policy area and find that elicitation of stakeholder preferences is important to preference-sensitive decisions as to understand the barriers towards screening attendance. In chapter 6 we presented a stated preference study where seven important attributes of lung screening policies were identified and weighted by a sample from the Dutch general population. Chapter 7 investigated how dependency relations in preference data can be modeled and demonstrated the advantages of using copula functions. In chapter 8, we compared different types of screening modalities and conclude they are valued differently by subgroups with different preferences or lung cancer risk. In the current chapter we will discuss the implications of the findings and suggest areas for further research.

Implications

A step toward evidence-based and patient-centered value frameworks

There is much discussion about involving patients in regulatory decisions and several initiatives are taken in which patients are invited to contribute to decision-making processes at FDA and EMA [1,2]. In addition, recent initiatives promote the development of

transparent evidence-based value frameworks for policy making allowing the formal input of different stakeholders [3,4] Most of these initiatives focus on how patient preferences can be measured, how patient representatives can be best involved or on how policy guidelines can ensure societal acceptance of decision-making processes [5,6]. An underdeveloped aspect of these initiatives is, however, how the perspective of patient populations relate to the available clinical evidence. In this thesis we have shown how preferences from large heterogeneous groups of patients can be used to weight an uncertain evidence base by developing a model based on stated preference studies and MCDA. A major advantage of working with the results of stated preference studies is that it is possible to look into the extent to which stochastic uncertainty and heterogeneity in preferences are present among stakeholders.

There are several examples in this thesis where these types of uncertainty directly influenced our confidence about which intervention was most valuable. First of all, in chapter 4 the value distributions of the various antiretroviral treatments had considerable overlap when stochastic uncertainty in preferences was included but most of this overlap was not present when only parameter uncertainty in clinical evidence was considered. Secondly, chapter 8 shows that although for most preference subgroups breath analysis has the highest mean value, this is not the case for the ‘sensitivity-specificity’ subgroup for who low-dose CT clearly has the highest value. Although it is insightful to know how a heterogeneous group of patients weight the evidence, this insight also challenges policy makers because they have to be clear on whose perspective on intervention value they will follow.

Although mathematical models cannot explain differences in preferences based on social, economic or ethical reasoning, they can provide insight into why people prefer some interventions over others and give decision makers a structured approach to search for ways to increase value for all stakeholders. For example, in chapter 8 we found that for most respondents low-dose CT had the lowest mean value although it has the best diagnostic accuracy backed up by a solid evidence base. Because our model provides insight into why this is the case, namely the respondents’ weights for non-clinical attributes and the low performance of CT on these attributes, we could provide concrete recommendations that may help policy makers to design a screening program that has the highest diagnostic accuracy but also provides the most value for the target population. However, policy makers will not always have the choice of designing interventions to suit the needs of most people. Therefore, policy makers will need to develop a value framework for evaluating interventions that integrates patient-based values with their own value

framework and concerns of other societal stakeholders in transparent decision-making processes [5,7,8].

A versatile approach for handling uncertainty in preferences and clinical evidence

Several well-developed best practice guidelines for MCDA in healthcare are available [9]. However, one of the yet unanswered questions is how uncertainty can be best taken into account. This thesis adapted a commonly used typology for uncertainty to the context of MCDA and identified five approaches for taking uncertainty into account for the context of healthcare [10]. Given the adoption of both deterministic and probabilistic approaches for uncertainty analyses in the context of Markov and micro-simulation models for cost-effectiveness analyses, it is still unknown which of the two is most appropriate for supporting researchers and analysts that apply MCDA for synthesizing evidence to inform regulatory decision making processes.

In this thesis we have demonstrated the versatility of the probabilistic approach for dealing with stochastic uncertainty, parameter uncertainty, and heterogeneity. We have shown that the different types of uncertainty found in the preferential and clinical evidence based can be incorporated into MCDA in an explicit and consistent manner. With the probabilistic approach, probability distributions can be assigned to uncertain parameters in the model based on best evidence. If there is very little evidence, uniform distributions can be used (see scenario analysis of chapter 3). Alternatively, existing data can be resampled using a bootstrap approach (chapter 3) or a multivariate normal distribution can be used to model correlations (chapter 4). Chapter 7 explored copula functions for modeling multivariate preference distributions. This approach seems especially promising because it allows for the modeling of correlations while analysts retain the ability to model marginal densities using different types of parametric distributions. Another part of the probabilistic approach that makes it attractive for analysts is Monte Carlo simulation. In chapters 3, 4, and 7 we have used Monte Carlo simulations to find the distribution of overall values that would have been very hard to derive analytically because they are a combination of a large number of probability distributions for preferences and clinical performance estimates. Overall, based on the findings in this thesis we argue that the probabilistic approach should be used by researchers and analysts when synthesizing uncertain evidence informing multiple model parameters.

Insight into the perceived value of screening programs

The area of lung cancer screening is at an important juncture. On the one hand, the evidence for the diagnostic accuracy of low-dose computed tomography (LDCT) is solidifying, with the influential NLST trial showing a 20% reduction in cancer-related mortality and two more major trials (NELSON and UKLS) presenting their results later in 2017. A disadvantage is that screening programs based on LDCT can be expensive and practical aspects of their service delivery may deter people from attending. On the other hand, innovative screening modalities such as circulating tumor cells or electronic noses promise cheaper screening at a more convenient location but the evidence for their diagnostic accuracy is still weak. The second part of this thesis has investigated this difficult policy issue for the context of the Netherlands. After having identified seven important attributes of screening, roughly one thousand respondents from the Dutch general population took part in a MCDA-based preference elicitation study. This study showed that location, mode, and sensitivity were the most important attributes. When we modeled the value of screening programs for this sample in chapter 8, we found that LDCT had the lowest value. This is surprising given its superior diagnostic accuracy but can be explained by the fact that a majority of respondents placed a high value on practical aspects of screening such as waiting time until results and location. Moving forward toward possible implementation of lung cancer screening, our results therefore urge policy makers to be aware of the preferences of the target population in order to maximize attendance.

Opportunities for future research

The studies in this thesis modeled how uncertainty in preferences and/or clinical evidence impacts the relative value and distribution of alternative interventions. Although we can calculate rank probabilities (e.g. the probability that a treatment has the highest rank and is thus preferred), we do not define a probability threshold at which decision makers can be confident that they made the right decision. Claxton argues that if decisions needs to be made now, the only rational choice would be to adopt the treatment with the highest expected value which, in our case, would mean the intervention with the highest probability of having the first rank [11]. If, however, the decision can be postponed to gather more evidence, it would be useful to know if such costly evidence gathering activities are likely to change the decision. There are several metrics that calculate the expected value of gathering more information (VOI) versus the expected value of deciding now. Although available and used in decision analytic modeling, VOI has to be explored in

MCDAs in a structured manner. We imagine that the value of information for parameter uncertainty in clinical evidence can be calculated in a similar way as for outcomes in conventional cohort models that are affected by this type of uncertainty [12]. For the stochastic uncertainty in preferences it may be useful to look at calculating the value of heterogeneity. This could be done by applying Basu's calculations on the Monte Carlo outcomes yielded by the models developed in this thesis: calculate the expected value when choosing the intervention with the highest value *in each simulation run*, and subtract from this the expected value when choosing the intervention that has the highest value *on average* [13]. In this way one can obtain an estimate of the value that is added for stakeholders by allowing multiple interventions to be used. With this knowledge, decision makers can then evaluate if they think that this added value outweighs the added cost or risks of using multiple interventions.

An outstanding issue for further research here is the comparison of values calculated with MCDA and (opportunity) costs. Because the values calculated with MCDA depend on its context-specific attributes, it is hard to compare value estimates between MCDA models. For example, we cannot make statements about the relative value of the antiretroviral treatments from chapter 4 compared to the screening programs in chapter 8 because they were calculated with structurally different models. If these values are incomparable due to different model structures, it will also not be possible to calculate a willingness-to-pay threshold that can be used to compare the results from different MCDA studies. Another issue related to uncertainty is that of how patients evaluate uncertain clinical performance. In this thesis we used the preferences of patients that were elicited by presenting respondents with fixed clinical estimates, while in reality clinical estimates are uncertain. Patients often decide not just on the mean clinical effect, but also on what they hope the effect will be. A clear example of this can be found in preferences for life saving cancer treatments where patients were found to focus on the maximum possible survival instead of the expected survival [14]. The main challenge for dealing with such effects is developing elicitation tools that are not biased by different interpretations of what uncertainty in a clinical outcomes means to them, while also being understandable for the general population that has, on average, a difficulty with interpreting probabilities [15].

Although we applied our model to a variety of cases, we did not involve policy makers. An important area for future research is therefore to assess and (if needed) to improve the match between our models and the needs of policy makers. A prominent barrier for this is the fact that most regulatory decision making processes are at this point qualitative, although initiatives are under way to investigate model-based quantitative decision making. The studies in this thesis cannot guide the decision-making process from start to finish, nor

do they provide guidance for the group decision-making process. However, our model is based on stated preference research and this can help decision makers include stakeholders in a structured and transparent way [16]. If a quantitative decision-making process is adopted that is founded on MCDA, further research will be needed to identify how the results from our stakeholder-centered MCDA models can be integrated in MCDA models that are constructed with policy makers to support the eventual group decision.

For the context of lung cancer screening there are also several areas for future research. The most prominent one at this point is that the evidence base for breath analysis and blood biomarkers should be expanded. It is important that randomized controlled trials are performed that compare these new screening modalities to low-dose CT. Given the fact that the diagnostic accuracy of LDCT is already very high, it can be worthwhile to explore methods that improve LDCT's performance on attributes other than sensitivity and specificity. Examples of this are the use of machine learning to reduce waiting time until results, or information campaigns to put the radiation burden of LDCT into perspective. This all should not be taken as to discount the value of novel screening modalities such as circulating tumor cells or electronic noses. These two modalities clearly have much potential to provide value for participants (as can be seen in chapter 8) if the diagnostic accuracy that is shown in the current studies can be replicated or improved upon in large randomized trials that compare the two novel modalities to LDCT.

Conclusion

The aim of this thesis was to investigate how the various types of uncertainty impact the outcomes of a stakeholder-based value framework for comparing medical interventions. We have shown that preferences elicited from a group of stakeholders with stated preference methods can be used to weight clinical evidence in an MCDA. Our models can help policy makers gain insight into how stakeholders value interventions as well as into which interventions have the highest value. Using the flexible probabilistic approach we developed an MCDA model that includes various types of uncertainty as well as dependency relations. Based on our findings we recommend that the probabilistic approach should be used by people who apply MCDA in the context of healthcare.

The second part of the thesis found that non-clinical attributes play an important part in how people value programs for lung cancer screening. It is shown that new screening modalities such as breath analysis and blood biomarkers may have a greater value for the target population than the established low-dose CT, primarily because of their better performance on non-clinical attributes. However, given that the evidence base for the

novel modalities is still weak, we argue that it is too soon to conclude that the novel modalities should be used for screening programs instead of low-dose CT. We argue that it would be better to consider our results in light of the question of how a screening program based on low-dose CT can be further improved and implemented by taking non-clinical factors into account.

Policy makers are increasingly challenged to make evidence-based decisions for a heterogeneous group of stakeholders given an uncertain evidence base. The work in this thesis develops quantitative tools with which policy makers can construct a structured and transparent value framework to help them address these difficult but important challenges.

References

- [1] Miller KL, Woodcock J. Value Assessment in the Regulatory Context. *Value Heal* 2017;20:296–8.
- [2] Ho MP, Gonzalez JM, Lerner HP, Neuland CY, Whang JM, McMurry-Heath M, et al. Incorporating patient-preference evidence into regulatory decision making. *Surg Endosc* 2015.
- [3] Bridges JFP, Jones C. Patient-based health technology assessment: a vision of the future. *Int J Technol Assess Health Care* 2007;23:30–5.
- [4] van Til JA, IJzerman MJ. Why Should Regulators Consider Using Patient Preferences in Benefit-risk Assessment? *Pharmacoeconomics* 2013:10–3.
- [5] Baltussen R, Paul M, Jansen M, Bijlmakers L, Grutters J, Kluytmans A, et al. Value Assessment Frameworks for HTA Agencies : The Organization of Evidence-Informed Deliberative Processes. *Value Heal* 2017;20:256–60.
- [6] Peretto EM, Oehrlein EM, Boutin M, Reid S, Gascho E. Value to Whom ? The Patient Voice in the Value Discussion. *Value Heal* 2017;20:286–91.
- [7] Rotter J, Foerster D, Bridges J. The changing role of economic evaluation in valuing medical technologies. *Expert Rev Pharmacoecon Outcomes Res* 2012;12:711–23.
- [8] Griffiths EA, Hendrich JK, Stoddart SDR, Walsh SCM. Acceptance of health technology assessment submissions with incremental cost-effectiveness ratios above the cost-effectiveness threshold. *Clin Outcomes Res* 2015;7:463–76.
- [9] Marsh K, IJzerman M, Thokala P, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making —Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Heal* 2016;19:1–13.
- [10] Briggs AH, Sculpher MJ, Buxton M. Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis. *Health Econ* 1994;3:95–104.
- [11] Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999;18:341–64.
- [12] Felli J, Hazen GB. Sensitivity analysis and the expected value of perfect information. *Med Decis Mak* 1998;18:95–109.
- [13] Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. *Med Decis Mak* 2009;27:112–27.

- [14] Shafrin J, Schwartz TT, Okoro T, Romley JA. Patient Versus Physician Valuation of Durable Survival Gains : Implications for Value Framework Assessments. *Value Heal* 2017;20:217–23.
- [15] Association AHC on HL for the C on A and AM. Health literacy: Report of the council on scientific affairs. *JAMA* 1999;281:552–7.
- [16] Oude Egbrink M, IJzerman M. The value of quantitative patient preferences in regulatory benefit-risk assessment. *J Mark Access Heal Policy* 2014;2:1–4.

Summary

In the **1st chapter** of this thesis we introduce that the aim of this thesis was to investigate how the various types of uncertainty influence the comparison of medical interventions in a value-framework incorporating stakeholder preferences. To introduce the rest of the work in the thesis this first chapter shortly describes the use of evidence for policy making, stated preference research, and multi-criteria decision analysis (MCDA).

Chapter 2 extended the typology of uncertainties developed by the ISPOR taskforce on good modeling practices to the context of MCDA. In the literature review of the same chapter, five approaches were identified to deal with these uncertainties, namely the deterministic, probabilistic, Bayesian, grey and fuzzy approaches. It was concluded that the deterministic approach was sufficient for most applications, but that more complex approaches were better suited for when multiple sources of uncertainty have to be considered simultaneously or when there are dependency relations. This is the case for the thesis' main aim of combining clinical performances and patient preferences, so the probabilistic approach was chosen to use in the rest of the thesis.

In **chapter 3**, it was applied to a case on antidepressants. Three hypothetical antidepressants were compared on four criteria. For each included treatment, the value distribution and rank probabilities were estimated. The impact of parameter uncertainty in clinical performances was larger than the impact of parameter uncertainty in patient preferences, and the assumption of uniform preference weights lead to substantially higher decision uncertainty. Scenario analyses clearly illustrated the different impact of parameter uncertainty in clinical performance and patient preferences.

The model was extended to include patient-specific preference variation in **chapter 4**. In this chapter, the value of six anti-retroviral treatments according to the preferences of HIV-positive persons was assessed. When only parameter uncertainty was considered there was a clear most valuable treatment, but when patient-specific preference variation was included there was still considerable uncertainty about the most valuable treatment.

Whereas the chapters in the first half of the thesis were not focused on a particular area of healthcare, we did focus on a specific application area in the second half. The chosen application area is lung cancer screening and the policy relevance of this area is explained

in **chapter 5**. There it is argued that the current low uptake of lung cancer screening as a national screening program is due to the large capacity impact of CT-based screening and preference-sensitive barriers to entry. A possible solution to these issues may be the introduction of new low-cost technologies such as breath analysis and biomarkers, and in developing screening policies using value frameworks that explicitly take into account the preferences of the eligible population.

Since to date no such preference studies have been done, in **chapter 6** a study like that is done. Based on interviews and a panel session, six important attributes of screening policies were identified. The subsequent MCDA-based elicitation found that the most important attributes according to the large and representative Dutch respondent sample were location of screening, mode of screening, and sensitivity. Cluster analyses revealed a distinction between subgroups focusing on clinical benefits/harms of screening, and subgroups focusing on practical aspects of the screening policy. The aim was to include the preferences from this study in a probabilistic MCDA model similar to those developed in chapters 3 and 4. Because several dependency relations were found in the preference data we could not assume (as was done in the previous chapters) that the weights distributions were independent.

Therefore, **chapter 7** investigated methods for fitting multivariate distributions for use in the probabilistic MCDA, focusing specifically on the use of the flexible copula functions approach. The study revealed the advantages of using copula functions for modeling multivariate preference distributions. The main advantage over conventional fitting methods lies in the fact that analysts can model the marginals and the dependency relations in two separate steps. This means the best fitting distribution can be chosen for each marginal, independently from any prior choice of (type of) multivariate distribution.

In the final **8th chapter**, probabilistic MCDA was applied to compare the value of screening policies based on CT, breath analysis, and biomarkers for subgroups defined on preferences or risk profile. To improve the value of LDCT-based screening for the eligible population we argue it is important to take non-clinical attributes of screening into account.

In final **9th chapter** we discuss our findings. We argue that our models can help policy makers gain insight into the value that stakeholders attach to interventions when the clinical evidence is uncertain. Future work should explore value of information methods and address the link to policy makers.

Nederlandse samenvatting

In het **hoofdstuk 1** van dit proefschrift introduceerden we dat het doel van dit proefschrift was om te onderzoeken hoe verschillende soorten onzekerheid de vergelijking tussen medische interventies beïnvloeden in een waarderingsraamwerk waar de voorkeuren van belanghebbenden (primair: patiënten) worden meegenomen. In dit hoofdstuk introduceerden we de rol van bewijs in besluitvorming, voorkeurenonderzoek, en multi-criteria beslissingsanalyse (in het engels multi-criteria decision analysis, afgekort MCDA)

In **hoofdstuk 2** breiden we de onzekerheidstypologie zoals die destijds geformuleerd is door de ISPOR werkgroep over best practices in modeleren uit naar de context van MCDA. Op basis van een systematische review konden we vijf methoden identificeren waarmee omgegaan kan worden met onzekerheid in MCDA: deterministische gevoeligheidsanalyse, probabilistische gevoeligheidsanalyse, Bayesiaanse methoden, 'grey' theorie en 'fuzzy set' theorie. We concludeerde dat de deterministische aanpak voldoende was voor de meeste toepassingen maar dat meer complexe methoden beter geschikt zijn wanneer er meerdere onzekerheidsbronnen zijn of wanneer er afhankelijkheidsrelaties in de data zitten. Deze twee zaken zijn het geval gezien het doel van dit proefschrift en we hebben ons in de rest van het proefschrift daarom gericht op de probabilistische methode.

In **hoofdstuk 3** passen we de probabilistische methode toe op een casus over drie hypothetische antidepressiva. Voor elk antidepressivum hebben we de waardeverdeling en de rangkansen geschat. Wat het meest waardevolle antidepressivum was hing meer af van parameteronzekerheid in klinische gegevens dan van parameteronzekerheid in patiëntvoorkeuren en was het hoogst als we aannamen dat de criteriumgewichten uniform verdeeld waren.

Het model is in **hoofdstuk 4** vervolgens uitgebreid om patiëntspecifieke variatie in voorkeuren voor antiretrovirale middelen mee te nemen. Wanneer alleen parameteronzekerheid mee werd genomen had één van de middelen duidelijk de hoogste waarde, maar er was onzekerheid over het meest waardevolle middel wanneer patiëntspecifieke variatie in voorkeuren mee werd genomen.

In de tweede helft van het proefschrift is specifiek bevolkingsonderzoek (screening) naar longkanker gekeken. In **hoofdstuk 5** leggen we uit waarom dit een actueel en relevant

onderzoeksveld is. We stellen dat de huidige lage opname van longkankerscreening (mede) komt door de hoge benodigde CT capaciteit en voorkeursgerelateerde factoren die mensen ervan zouden weerhouden deel te nemen. Een oplossing voor deze problemen zou zijn om gebruik te maken van nieuwe screeningstechnologieën zoals ademanalyse en biomarkers en om bij het ontwikkelen van bevolkingsonderzoek naar longkanker expliciet rekening te houden met de voorkeuren van potentiële deelnemers.

Helaas zijn er geen onderzoeken bekend die de voorkeuren van potentiële deelnemers onder de loep nemen. Om belangrijke beslissingscriteria te ontdekken hebben we daarom in **hoofdstuk 6** interviews en een panel sessie georganiseerd. Vervolgens hebben we een MCDA vragenlijst gemaakt om de criteriumgewichten te kunnen meten. Na de vragenlijst uitgezet te hebben onder een groep van ongeveer duizend Nederlanders vonden we dat de drie belangrijkste criteria de volgende waren: locatie, modus (beleving van onderzoek), en sensitiviteit. We zagen een splitsing tussen klinische georiënteerde subgroepen in voorkeuren aan de ene kant en praktisch georiënteerde subgroepen aan de andere kant.

Omdat we afhankelijkheidsrelaties (correlaties) vonden in de voorkeuren data, hebben we in **hoofdstuk 7** methoden vergeleken waarmee multivariate kansverdelingen geschat kunnen worden. We concludeerden dat de flexibele copula verdeling geschikt was voor onze toepassing. Het voordeel van een copula is dat de marginalen en de afhankelijkheidsrelaties in twee aparte stappen geschat kunnen worden. Dit betekent dat de best passende verdeling gekozen kan worden per criterium, in plaats van vast te zitten aan een bepaalde univariate verdeling die hoort bij de vooraf gekozen multivariate verdeling.

In het **8e hoofdstuk** hebben we probabilistische MCDA gebruikt om CT, ademanalyse en biomarkers als technologie voor bevolkingsonderzoek naar longkanker te vergelijken. We vergeleken de waarde hiervan voor subgroepen in voorkeuren en in risicoprofiel. Om de waarde van CT voor deelnemers te verhogen is het belangrijk om te kijken naar niet-klinische factoren.

In **hoofdstuk 9** bespreken we onze bevindingen. We stellen dat onze modellen beleid kunnen ondersteunen door inzicht te geven in hoe belanghebbenden medische interventies waarderen als de klinische gegevens onzeker zijn. Toekomstig onderzoek zou zich moeten richten op value of information methoden en op de link met beleidmakers.

Short CV of author Henk Broekhuizen

Professional experience

January 2017 – current	<i>Postdoc, dept. Health Evidence at Radboud University Medical Centre</i>
September 2016 – current	<i>Data scientist / R specialist, X-IS</i>
October 2015 – December 2015	<i>Visiting researcher, Centre for Health Economics at University of York</i>
September 2012 – June 2017	<i>PhD student, dept. Health Technology and Services Research at Twente University</i>

Journal articles

Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber AB, Jansen JP, IJzerman MJ. *Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis*. BMC Medical Informatics and Decision Making 2015;15:1–10

Broekhuizen H, Groothuis-Oudshoorn CGM, van Til JA, Hummel JM, IJzerman MJ. *A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions*. Pharmacoeconomics 2015;33:445–55.

Broekhuizen H, IJzerman M, Hauber AB, Groothuis-Oudshoorn CGM. *Weighing clinical evidence using patient preferences: an application of probabilistic multi-criteria decision analysis*. Pharmacoeconomics 2017;35:259–69.

Broekhuizen H, Groothuis-Oudshoorn CGM, Vliegthart R, Groen H, IJzerman MJ. *Public preferences for lung cancer screening policies*. Accepted for publication in Value in Health.

Book chapters

Groothuis-Oudshoorn CGM, Broekhuizen H, van Til J. *Dealing with uncertainty in the analysis and reporting of MCDA*. In: Marsh K, Goetghebeur M, Thokala P, Baltussen R, editors. Multi-Criteria Decision Analysis to Support Healthcare Decisions. 1st ed., Springer; 2017, p. 65–83. doi:10.1007/978-3-319-47540-0.

Conference presentations

Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber AB, IJzerman MJ.

Integrating patient preferences and clinical trial data in a Bayesian model for benefit-risk assessment. At the 34th annual meeting for the Society for Medical Decision Making (2012) in Phoenix, AZ, US and at the 15th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2012) in Berlin, Germany.

Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber AB, IJzerman MJ.

Integrating patient preferences and clinical trial data in a quantitative model for benefit-risk assessment. At the 25th annual eurometing of the Drug Information Association (2013) in Amsterdam, the Netherlands.

Groothuis-Oudshoorn CGM, Broekhuizen H, Hummel JM, van Til JA, IJzerman MJ.

Parameter uncertainty in value based multi criteria decision analysis: a systematic review of methods. At the 16th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2013) in Dublin, Ireland and at the 15th biennial European meeting of the Society for Medical Decision Making (2013) in Antwerp, Belgium.

Broekhuizen H, Groothuis-Oudshoorn CGM, Hauber AB, IJzerman MJ. *A probabilistic framework to combine discrete choice preferences and uncertain clinical data.* At the 15th biennial European meeting of the Society for Medical Decision Making (2013) in Antwerp, Belgium.

Broekhuizen H, IJzerman MJ, Hauber AB, Groothuis-Oudshoorn CGM.

Patient preferences and HIV drugs: what about uncertainty? At the 17th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2014) in Amsterdam, the Netherlands.

IJzerman MJ, Marsh K, Broekhuizen H, Groothuis-Oudshoorn CGM.

Workshop: supporting decision making with MCDA: recommendations for dealing with uncertainty. At the 17th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2014) in Amsterdam, the Netherlands.

Broekhuizen H, Hummel JM, Groothuis-Oudshoorn CGM, IJzerman MJ.

Comparing patient preferences for medical treatments with PROMETHEE II: a pilot study. At the 2nd International MCDA workshop on PROMETHEE (2015) in Brussels, Belgium.

Broekhuizen H, IJzerman MJ, Hauber AB, Groothuis-Oudshoorn CGM.

Building a bridge between clinical evidence research and patient preference research. At the 17th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2014) in Amsterdam, the Netherlands.

Schoutrop A, Broekhuizen H, Siesling S, IJzerman MJ.

The consequences of replacing the Fleischner guidelines by a software-based volume doubling time technique: an early-stage research. At the 18th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2015) in Milan, Italy.

Broekhuizen H, Broekhuizen MA, IJzerman MJ.

Using geographic information systems and multi-criteria decision analysis to address spatially explicit problems in healthcare policy decision making. At the 18th annual European congress of the International Society for Pharmacoeconomics and Outcomes Research (2015) in Milan, Italy.

Hummel JM, Broekhuizen H, Siegersma S, de Geus-Oei LF.

Early assessment of the added value of FDG-PET/CT and MRI-DWI in monitoring treatment response of stage III non-small cell lung cancer. At the 14th mediterranean conference on medical and biological engineering and computing (2016) in Paphos, Cyprus.

IJzerman MJ, Broekhuizen H, Groothuis-Oudshoorn CGM, Vliegenthart R, Groen H.

Elicitation of public preferences for lung cancer screening using three screening modalities. At the annual meeting of the European Society for Medical Oncology (2016) in Copenhagen, Denmark.

Students supervised

Lea Rietkötter (MSc Health Sciences, 2014). *Ending the war in multi-criteria decision analysis: Taking the best from two worlds : the development and evaluation of guidelines for the use of MACBETH in multi-criteria group decision making for the assessment of new medical products.* Supervised with Marjan Hummel and Karin Groothuis

Klaske Siegersma (MSc Health Sciences, 2015). *Assessment of the added value and uncertainty of implementation of FDG-PET/CT and MRI-DWI in monitoring treatment response of stage III non-small cell lung cancer*. Supervised with Marjan Hummel and Lioe-Fee de Geus-Oei.

Dankwoord

De afgelopen 4+ jaar was een lange en verschillend gekleurde tijd maar ik ben blij met het resultaat. Ik ben niet zo van de sentimentele verhalen maar uiteraard zijn er een hoop mensen die ik wil bedanken voor hun hulp afgelopen jaren. Mocht ik iemand per abuis niet noemen (vergeten is mijn specialiteit...), dan excuus. Vier jaar is een lange tijd en er zijn in die tijd een heleboel mensen op mijn pad langs gekomen.

Allereerst wil ik graag Maarten bedanken voor de geboden kans en het getoonde vertrouwen. Ik vond het erg uitdagend en leuk om aan een project te werken waar van tevoren niet vast stond wat het eindpunt zou zijn. Toen ik in 2012 tijdens een werkcollege jou aansprak om te zeggen dat ik modelleren wel 'cool' vond en me afvroeg of hier ook afstudeeropdrachten over waren, had ik niet kunnen bedenken dat het uiteindelijk tot zo'n mooi boekje zou leiden. Ik heb een heleboel mogen leren de afgelopen tijd en heb een hoop interessante plekken mogen bezoeken. Bedankt dat je me steeds hebt uitgedaagd en fijn dat je om wist te gaan met mijn eigenwijsheid zo af en toe.

Karin, het was erg fijn om jou als dagelijks begeleider te hebben. Ik had het idee dat je niet alleen mijn werk wilde ondersteunen, maar ook geïnteresseerd was in de ontwikkeling van mij als persoon. Jouw perfectionisme en wil om te zorgen dat alles klopte was opbouwend geformuleerd, maar frustreerde me soms. Ik had dat echter absoluut nodig. In mijn verdere werkende leven zal ik altijd een 'kleine Karin' in m'n hoofd (nodig) hebben die voorkomt dat ik te snel tevreden ben met een model of paper, die zorgt dat ik dingen goed doordenk en goede vragen stel.

Een proefschrift schrijf je niet alleen met je begeleiders, daarom wil ik ook graag enkele co-auteurs bedanken. Brett, thank you for working with me on the third and fourth chapter. It was a pleasure to discuss the models with an expert and your feedback on my papers was always constructive and insightful. My time at CHE in York has been great and really educational. Andrea, although copulas were a difficult topic I am proud we figured it out and were able to write a nice chapter on its application. Thanks for this opportunity. Thanks also to Jeroen and Sebastian for working with me on the third and fifth chapter, respectively. Janine, bedankt voor met meedenken en schrijven van de review en de verscheidene goede inhoudelijke gesprekken over MCDA. Graag wil ik Rozemarijn en Harry bedanken voor hun bijdragen aan het tweede deel van het proefschrift. Het was voor iemand met een methodologische achtergrond van onschatbare waarde om co-auteurs te

hebben die met een klinische blik naar het onderzoek keken. Hopelijk kan ons gezamenlijke werk bijdragen aan de discussie omtrent longkankerscreening. Marjan, bedankt dat ik mee mocht doen met jouw onderzoekslijn naar AHP/MACBETH in panelsessies voor early HTA toepassingen. Ik heb het erg leerzaam en leuk gevonden. Volgens mij hebben we twee mooie afstudeerverslagen af mogen leveren. Voor hoofdstuk zes heb ik mensen geïnterviewd en een panelsessie gehouden. Bedankt aan Klaske voor technische ondersteuning tijdens de panelsessie en aan alle deelnemers.

Members of the committee, thank you for taking the time for my work in your no doubt very busy schedules.

Jan, Wouter, Rob, en Leon wil ik graag bedanken voor hun flexibiliteit in de afgelopen periode dat ik al wel voor hen werkte maar ook nog met de laatste dingen van de promotie bezig was. Ik zie er naar uit om verder met jullie samen te werken.

Collega's van HTSR, bedankt voor de goede tijd op de afdeling. Leuk om zo een tijd samen op te trekken als collega's (en/of lotgenoten). Alhoewel onze onderwerpen soms heel anders zijn heb ik veel gehad aan de gesprekken en gezellige momenten. Alle namen noemen hier gaat te ver, maar een speciaal bedankje voor Rolf, Sandra, Annemieke, Angarath, Marieke, Michelle, en Sarah die mij als kamergenoot toelieten (ondanks dat ik zoveel snoep).

Björn en Tjark, bedankt voor de gezellige avonden en de discussies die zowel interessant als vermakelijk waren. Fijn dat jullie me tijdens de verdediging als paranimfen bij willen staan! Andere familie en vrienden wil ik bedanken voor het luisteren naar mijn soms lastig te volgen werkverhalen en voor het geduld als ik soms veel bezig moest zijn met werken. Dat laatste geldt vooral voor Miranda; bedankt dat je me de kans en de tijd hebt gegund om te promoveren in het verre Enschede. Ik ben benieuwd waar de toekomst ons gaat brengen.

Vader in de hemel, bedankt voor uw trouw. Er is in vier jaar veel gebeurd, bedankt dat U erbij was.

U wijst mij een weg, als ik zelf geen uitkomst zie.

Langs wegen die geen mens bedenkt, maakt U mij uw wil bekend.

U geeft elke dag, nieuwe liefde, nieuwe kracht.

Als ik mijn hand in uw hand leg, wijst U mij de weg.

Getekend op 19-5-2017, Henk