

INCORPORATING SOCIETAL PREFERENCES IN REIMBURSEMENT DECISIONS

RELATIVE IMPORTANCE OF DECISION CRITERIA ACCORDING TO BELGIAN CITIZENS



INCORPORATING SOCIETAL PREFERENCES IN REIMBURSEMENT DECISIONS

RELATIVE IMPORTANCE OF DECISION CRITERIA ACCORDING TO BELGIAN CITIZENS

IRINA CLEEMPUT, STEPHAN DEVRIESE, LAURENCE KOHN, CARL DEVOS, JANINE VAN TIL, KARIN GROOTHUIS-ODSHOORN, PIETER VANDEKERCKHOVE, CARINE VAN DE VOORDE



Title:	Incorporating societal preferences in reimbursement decisions – Relative importance of decision criteria according to Belgian citizens
Authors:	Irina Cleemput (KCE), Stephan Devriese (KCE), Laurence Kohn (KCE), Carl Devos (KCE), Janine van Til (Universiteit Twente, the Netherlands), Karin Groothuis-Oudshoorn (Universiteit Twente, the Netherlands), Pieter Vandekerckhove (Mapi, United Kingdom), Carine Van de Voorde (KCE)
Project coordinator:	Nathalie Swartenbroekx (KCE)
External experts:	Lieven Annemans (Ugent), Jaak Billiet (KU Leuven), Ri De Ridder (RIZIV – INAMI), Myriam De Spiegelaere (ULB), Brigitte Duvieusart (Koning Boudewijnstichting – Fondation Roi Baudouin), Margreet Franken (Erasmus Universiteit Rotterdam, the Netherlands), Johan Hansen (Nederlands Instituut voor Onderzoek van de Gezondheidszorg (NIVEL), the Netherlands), Mark Koopmanschap (Erasmus Universiteit Rotterdam, the Netherlands), Erik Schokkaert (KU Leuven), Martina Vandebroek (KU Leuven), Tinne Vandensande (Koning Boudewijnstichting – Fondation Roi Baudouin), Toon Vandeveldde (KU Leuven), Bert Winnen (RIZIV – INAMI)
External validators:	Ann Carton (Studiedienst van de Vlaamse Regering), Roselinde Kessels (Universiteit Antwerpen), Aki Tsuchiya (University of Sheffield, United Kingdom)
Acknowledgements:	Stefan Van De Venster (FOD Binnenlandse zaken – SPF Intérieur), Peter Raeymaekers (LyRaGen), Gerrit Rauws (Koning Boudewijnstichting – Fondation Roi Baudouin), Monique Bosson (Belgisch Raadgevend Comité voor Bio-ethiek – Comité Consultatif de Bioéthique de Belgique), Johan Van der Heyden (WIV – ISP), Stefaan Demarest (WIV - ISP), Sylviane de Viron (ISP – WIV), Edwin Pelfrene (SVR), Ann Carton (SVR), Marc Callens (SVR)
Other reported interests:	<p>Fees or other compensation for writing a publication or participating in its development: Peter Raeymaekers (freelance collaborator / consultant Koning Boudewijnstichting – Fondation Roi Baudouin)</p> <p>Participation in scientific or experimental research as an initiator, principal investigator or researcher: Roselinde Kessels (co-researcher / co-author doctoral research Jeroen Luyten 'Public preferences for prioritizing preventive and curative health care interventions: a discrete choice experiment')</p> <p>Presidency or accountable function within an institution, association, department or other entity on which the results of this report could have an impact: Ri De Ridder (CEO RIZIV – INAMI), Bert Winnen (Head Medical Department RIZIV – INAMI)</p>
Layout:	Ine Verhulst

**Disclaimer:**

- The external experts were consulted about a (preliminary) version of the scientific report. Their comments were discussed during meetings. They did not co-author the scientific report and did not necessarily agree with its content.
- Subsequently, a (final) version was submitted to the validators. The validation of the report results from a consensus or a voting process between the validators. The validators did not co-author the scientific report and did not necessarily all three agree with its content.
- Finally, this report has been approved by common assent by the Executive Board.
- Only the KCE is responsible for errors or omissions that could persist. The policy recommendations are also under the full responsibility of the KCE.

Publication date: 22 December 2014 (2nd print; 1st print: 22 December 2014)

Domain: Health Services Research (HSR)

MeSH: Decision Making; Health Insurance Reimbursement; Consumer participation

NLM Classification: W 84.3 Research (General)

Language: English

Format: Adobe® PDF™ (A4)

Legal depot: D/2014/10.273/91

Copyright: KCE reports are published under a “by/nc/nd” Creative Commons Licence
<http://kce.fgov.be/content/about-copyrights-for-kce-reports>.

**How to refer to this document?**

Cleemput I, Devriese S, Kohn L, Devos C, van Til J, Groothuis-Oudshoorn K, Vandekerckhove P, Van de Voorde C. Incorporating societal preferences in reimbursement decisions – Relative importance of decision criteria according to Belgian citizens. Health Services Research (HSR) Brussels: Belgian Health Care Knowledge Centre (KCE). 2014. KCE Reports 234. D/2014/10.273/91.

This document is available on the website of the Belgian Health Care Knowledge Centre.



■ TABLE OF CONTENTS

LIST OF FIGURES.....	4
LIST OF TABLES	7
LIST OF ABBREVIATIONS.....	10
■ ABSTRACT	11
■ SUMMARY	13
■ SCIENTIFIC REPORT	25
1 BACKGROUND AND SCOPE	25
1.1 SCOPE OF THE STUDY	25
1.2 LEGITIMATE DECISION-MAKING.....	26
1.2.1 Accountability for reasonableness.....	26
1.2.2 Putting accountability for reasonableness into practice	26
1.2.3 Accountability for reasonableness in a deliberative decision-making system.....	27
1.2.4 Preferences and values	28
1.3 MULTI-CRITERIA DECISION ANALYSIS (MCDA)	28
1.3.1 Structuring the debate and making decision processes transparent	28
1.3.2 A five-question framework	29
1.3.3 MCDA in practice.....	31
2 OBJECTIVES	35
3 GENERAL METHODOLOGICAL APPROACH	36
4 LITERATURE REVIEW	36
4.1 METHODS	36
4.1.1 Literature search strategy.....	36
4.1.2 Scope of the literature review	37
4.1.3 Selection of studies	37
4.1.4 Data extraction methodology.....	38
4.1.5 External expert group discussions	40
4.2 RESULTS: CRITERIA FOR MAKING REIMBURSEMENT DECISIONS	41



4.2.1	Flow-chart literature review	41
4.2.2	Classification of principles for rational resource use and choice criteria.....	42
4.2.3	General principles for resource allocation	42
4.2.4	Patient-related criteria	46
4.2.5	Condition-related criteria	48
4.2.6	Intervention-related criteria.....	50
4.3	RESULTS: PREFERENCE ELICITATION TECHNIQUES.....	54
4.3.1	Assumptions about human behaviour	54
4.3.2	Techniques for measuring preferences.....	59
4.3.3	Evaluation of preference elicitation techniques.....	65
4.3.4	Discussion and conclusion	69
5	SURVEY ON THE RELATIVE IMPORTANCE OF DIFFERENT DECISION CRITERIA FOR REIMBURSEMENT	71
5.1	METHODS.....	71
5.1.1	Choice of data collection technique.....	71
5.1.2	Questionnaire development process.....	72
5.1.3	Sample selection	73
5.1.4	Anonymity	74
5.1.5	Invitation letters and reminders	76
5.1.6	Preference elicitation technique	76
5.1.7	Criteria	77
5.1.8	Discrete choice experiment choice sets.....	80
5.1.9	Design of DCE choice sets.....	83
5.1.10	Survey versions.....	88
5.1.11	Survey completion and response registration process	88
5.1.12	Survey structure	89
5.1.13	Data analysis	89
5.2	TEST-RETEST RELIABILITY	93
5.3	SAMPLE CHARACTERISTICS	94
5.3.1	Demographics of the general population sample.....	94



5.3.2	Demographics of the decision makers' sample.....	97
5.3.3	Comparison of the general population and decision makers' sample.....	99
5.3.4	Response by reminder	105
5.3.5	Differences in population sample characteristics by reminders	107
5.3.6	Reflections on the sample characteristics	113
5.4	CHOICE SET ANALYSIS: TOTAL SAMPLE	114
5.4.1	Reported certainty of choices	114
5.4.2	Attribute weights in Therapeutic need domain	117
5.4.3	Attribute weights in Societal need domain	130
5.4.4	Attribute weights in Added value domain	136
5.5	COMPARISON OF THE WEIGHTS OF SUBSAMPLES OF THE GENERAL POPULATION.....	148
5.5.1	Weights by subgroup defined by number of reminders received	148
5.5.2	Weights by subgroup defined by age	151
5.5.3	Weights by subgroup defined by self-reported health status	154
5.5.4	Weights by subgroup defined by uncertainty	157
6	FUTURE USE THE RESULTS OF THIS STUDY	161
6.1	PRACTICAL STEPS IN THE MCDA	161
6.2	CONSIDERATIONS BEYOND THE MCDA	165
6.3	SCORING RULES	166
7	GENERAL DISCUSSION	168
7.1	ISSUES IN MCDA RESEARCH	168
7.1.1	Choice of the MCDA approach.....	168
7.1.2	Double counting criteria.....	169
7.1.3	Gaps in evidence and bias	169
7.1.4	Inter-rater consistency	169
7.1.5	Choice of the weighting technique	169
7.1.6	Uncertainty in evidence	170
7.1.7	Interpretation of MCDA scores	171
7.1.8	Impact of MCDA	172
7.2	ASSUMPTIONS AND LIMITATIONS	172



7.2.1	Belgian citizens are not mere QALY maximizers	172
7.2.2	A multi-layer MCDA is more manageable and acceptable to policymakers than an all-in-one MCDA	172
7.2.3	A reasonableness test of the rankings will have to be performed.....	173
7.3	WHO SHOULD BE INVOLVED IN THE APPLICATION OF THE MCDA TOOL?	173
8	CONCLUSION	175
■	REFERENCES	176

LIST OF FIGURES

Figure 1 – Preparedness to pay (more) for a new intervention	23
Figure 1 – Model of accountability for priority setting in health care	27
Figure 2 – Steps to be taken in the development of an MCDA framework.....	29
Figure 3 – Multi-criteria decision analysis: how it could work in practice	31
Figure 4 – Flow chart literature search.....	41
Figure 5 – Changing marginal rate of substitution between attributes.....	55
Figure 6 – Example of a choice question	61
Figure 7 – Example of DCE question in Mortimer et al (2008)	62
Figure 8 – Example of a question using pie method	63
Figure 9 – Example of a general best-worst scaling exercise – case 2.....	63
Figure 10 – Example of a best-worst scaling exercise – case 3.....	64
Figure 11 – Phase 1 of the analytical hierarchy process: paired comparison between criteria.....	65
Figure 12 – Survey process	75
Figure 13 – Blocks, attributes and levels used in the survey	78
Figure 14 – Example of a DCE question for therapeutic need	81
Figure 15 – Example of a DCE question for societal need	81
Figure 16 – Example of a DCE question for added value of treatment	82
Figure 17 – Choice set design in the Societal need domain.....	84
Figure 18 – Choice set design Therapeutic need domain	85
Figure 19 – Choice set design Added value domain	87
Figure 20 – Survey versions	88
Figure 21 – Distribution of time between test and retest.....	93



Figure 22 – Correspondence between answers in test and retest	94
Figure 23 – Age and gender distribution of the general population analysis sample (complete) compared to the respondents who didn't complete all choice sets (not complete).....	95
Figure 24 – Age and gender distribution of the general population sample compared to the Belgian population	97
Figure 25 – Age and gender distribution of the decision makers' sample	98
Figure 26 – Distribution of educational levels in the study sample	100
Figure 27 – Self-reported health status.....	101
Figure 28 – Self-reported health status in the general population sample, compared to Health Interview Survey 2013	102
Figure 29 – Affordability of health care	103
Figure 30 – Respondents' living conditions.....	104
Figure 31 – Reception of responses over time	105
Figure 32 – Response by medium in function of number of reminders	106
Figure 33 – Time of completion by number of reminders	106
Figure 34 – Proportion of questionnaires returned after initial invitation, one, two or three reminders, by survey version and response medium.....	107
Figure 35 – Number of reminders by age and gender	108
Figure 36 – Number of reminders by family living conditions	109
Figure 37 – Number of reminders by professional status	110
Figure 38 – Number of reminders by educational level	111
Figure 39 – Number of reminders by self-reported health status	112
Figure 40 – Number of reminders by perceived affordability of health care	113
Figure 41 – Reported uncertainty of choices per domain	115
Figure 42 – Reported uncertainty of choices per domain	115
Figure 43 – Reported uncertainty of choices per domain and number of reminders.....	116
Figure 44 – Probabilities of choosing a scenario as having a higher therapeutic need out of the full set of scenarios, general population	121
Figure 45 – Probabilities of choosing a scenario as having a higher therapeutic need out of the full set of scenarios, decision makers.....	125
Figure 46 – Comparison of relative preference weights by method in the Therapeutic need domain, general population sample	129



Figure 47 – Comparison of relative preference weights by method in the Therapeutic need domain, decision maker sample.....	129
Figure 48 – Probabilities of choosing a scenario as having a higher societal need out of the full set of scenarios, general population	131
Figure 49 – Probabilities of choosing a scenario as having a higher societal need out of the full set of scenarios, decision makers.....	133
Figure 50 – Comparison of relative preference weights by method in the Societal need domain, general population sample	135
Figure 51 – Comparison of relative preference weights by method in the Societal need domain, decision maker sample.....	135
Figure 52 – Probabilities of choosing a scenario as having a higher added value out of the full set of scenarios, general population.....	138
Figure 53 – Probabilities of choosing a scenario as having a higher added value out of the full set of scenarios, decision makers.....	144
Figure 54 – Comparison of relative preference weights by method in the Added value domain, general population sample	146
Figure 55 – Comparison of relative preference weights by method in the Added value domain, decision maker sample.....	146
Figure 56 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of number of reminders received	149
Figure 57 – Relative weights of decision criteria for societal need by subgroup defined in function of number of reminders received	150
Figure 58 – Relative weights of decision criteria for added value by subgroup defined in function of number of reminders received	151
Figure 59 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of age	152
Figure 60 – Relative weights of decision criteria for societal need by subgroup defined in function of age.....	153
Figure 61 – Relative weights of decision criteria for added value by subgroup defined in function of age	154
Figure 62 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of self-reported health status.....	155
Figure 63 – Relative weights of decision criteria for societal need by subgroup defined in function of self-reported health status.....	155
Figure 64 – Relative weights of decision criteria for added value need by subgroup defined in function of self-reported health status.....	157



Figure 65 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of certainty of responses	158
Figure 66 – Relative weights of decision criteria for societal need by subgroup defined in function of certainty of responses	159
Figure 67 – Relative weights of decision criteria for added value by subgroup defined in function of certainty of responses	160
Figure 68 – Preparedness to pay (more) for a new intervention	164

LIST OF TABLES

Table 1 – Criteria included in the survey	15
Table 2 – Weights for criteria in the Therapeutic Need domain	18
Table 3 – Weights for criteria in the Societal need domain	18
Table 4 – Weights for criteria determining the added value of new treatments	19
Table 5 – Weights of the criteria in each cluster	21
Table 1 – Key questions and possible criteria for a drug reimbursement appraisal process (MCDA framework)	30
Table 2 – General guidance for the application of MCDA and application to reimbursement decision making in health care	33
Table 3 – Search strings	36
Table 4 – Inclusion and exclusion criteria for the literature review	38
Table 5 – Overview of strengths, weaknesses, opportunities and threats of different behavioural assumptions	58
Table 6 – Overview of strengths, weaknesses and limitations of different preference elicitation techniques	70
Table 7 – Gender and language distribution of test-retest sample	93
Table 8 – Language distribution	96
Table 9 – Response rate per decision-maker organisation	99
Table 10 – Actual and predicted percentage of choice for each alternative	117
Table 11 – Therapeutic need: goodness of fit statistics	117
Table 12 – Therapeutic need: model summary for the general population sample	118
Table 13 – Some examples of conditions with their level of therapeutic need according to the model	120
Table 14 – Differences in coefficients between attribute levels	123
Table 15 – Therapeutic need: model summary for the decision maker sample	124
Table 16 – Log-likelihood of models in the Therapeutic need domain, general population sample	126



Table 17 – Log-likelihood of models in the Therapeutic need domain, decision maker sample	126
Table 23 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, general population sample	127
Table 24 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, decision maker sample.....	127
Table 25 – Coefficient range weights in the Therapeutic need domain, general population sample	127
Table 26 – Coefficient range weights in the Therapeutic need domain, decision maker sample.....	128
Table 27 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, general population sample (coefficient range method)	128
Table 28 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, decision maker sample (coefficient range method)	128
Table 24 – Actual and predicted percentage of choice for each alternative	130
Table 25 – Societal need: goodness of fit statistics	130
Table 26 – Societal need: model summary for the general population sample	130
Table 27 – Some examples of conditions with their level of societal need according to the general public	131
Table 28 – Differences in coefficients between attribute levels	131
Table 29 – Societal need: model summary for the decision maker sample	132
Table 30 – Log-likelihood of models in the Societal need domain, general population sample	133
Table 31 – Log-likelihood of models in the Societal need domain, decision makers sample.....	134
Table 32 – Derivation of the weights for a priori selected criteria in the Societal need domain, general population sample (coefficient range method)	134
Table 33 – Derivation of the weights for a priori selected criteria in the Societal need domain, decision maker sample (coefficient range method).....	134
Table 34 – Actual and predicted percentage of choice for each alternative	136
Table 35 – Societal need: goodness of fit statistics	136
Table 36 – Added value: model summary for the general population sample.....	137
Table 37 – Differences in coefficients between attribute levels	139
Table 38 – Added value: model summary for the decision maker sample	141
Table 39 – Log-likelihood of models in the Added Value domain, general population sample	142
Table 40 – Log-likelihood of models in the Added value domain, decision makers sample.....	142



Table 41 – Derivation of the weights for a priori selected criteria in the Added value domain, general population sample (coefficient range method)	145
Table 42 – Derivation of the weights for a priori selected criteria in the Added value domain, decision maker sample (coefficient range method).....	145
Table 43 –Scoring table	161
Table 44 – Weights of decision criteria per domain as measured in the general public.....	163
Table 45 – Scoring rules for the criteria	167



LIST OF ABBREVIATIONS

ABBREVIATION	DEFINITION
AHP	Analytical Hierarchy Process
CTIIMH – CRIDMI	Committee for the Reimbursement of Invasive Medical Devices and Implants ('Commissie Tegemoetkoming Implantaten en Invasieve Medische Hulpmiddelen'/'La commission de remboursement des implants et des dispositifs médicaux invasifs')
CTG – CRM	Drug Reimbursement Committee ('Commissie voor Tegemoetkoming Geneesmiddelen'/'Commission de Remboursement des Médicaments')
DCE	Discrete Choice Experiment
EQ-5D-5L	EuroQol health-related quality of life instrument with five dimensions and five levels per dimension
FAGG – AFMPS	Federal Agency for Medicines and Health Products ('Federaal Agentschap voor Geneesmiddelen en Gezondheidsproducten'/'Agence Fédérale des Médicaments et des Produits de Santé')
HRQoL	Health-related Quality of Life
IQR	Inter-quartile range
MCDA	Multi-Criteria Decision Analysis
NICE	National Institute for Health and Care Excellence
PET	Positron Emission Tomography
QALYs	Quality-Adjusted Life Years
RIZIV – INAMI	National Institute for Health and Disability Insurance ('Rijksinstituut voor Ziekte- en Invaliditeitsverzekering'/'Institut National d'Assurance Maladie-Invalidité')



■ ABSTRACT

BACKGROUND

Legitimate health care reimbursement decisions should take social preferences into account. However, decision makers have very little information about the preferences of the general public with respect to reimbursement criteria. Criteria are multiple, which complicates decision-making processes. Multi-criteria decision analyses (MCDA) can help to structure the process and make it more consistent, but requires criteria weights and criteria scores.

OBJECTIVE

This study aims at measuring public preference weights for reimbursement criteria. Guidance for scoring the criteria is outside the scope of the current study.

METHODS

The study follows a previously developed hierarchical decision framework, which presumes that an intervention can only be worthwhile reimbursing if there is a need for such an intervention and the added value of the intervention is sufficient. A large survey was performed in the general public and decision makers, using discrete choice experiments (DCEs). Consistent with the framework, the survey was composed of three blocks:

- (1) therapeutic need, i.e. the need for a better treatment in a particular disease given the treatment already available, as determined by the quality of life under current treatment, the impact of the disease on life expectancy despite current treatment and the current treatment's inconvenience;
- (2) societal need, as determined by the prevalence of the disease and the public expenditures per patient with that disease;
- (3) added value of a new intervention relative to the best alternative intervention, as determined by the impact of that new intervention on all previous criteria.



Responses of 4288 participants from the general public (out of 20 000 people invited) and 161 participants (38.2%) from the decision makers were used for analysis. A multinomial logistic regression analysis was performed to analyse the data. Level-independent criteria weights were determined using two different methods to test the robustness of the weights.

RESULTS

In the appraisal of therapeutic need, the general public as well as the policy makers gave the highest weight to the current quality of life. For the general public, the inconvenience of current treatment is more important than the impact of the disease on life expectancy, despite current treatment. This is the other way around for the decision makers.

In the appraisal of societal need, people from the general public give more weight to the impact of a disease on public expenditures than to the prevalence of the disease, unlike the decision makers. The order of the weights differed between the weight-determination methods, however, so no clear conclusions can be drawn.

In the appraisal of the added value of new interventions, the general public gives the highest weight to the intervention's impact on quality of life, followed by its impact on the prevalence of the disease and on life expectancy. Decision makers have the same preference order, but their weight for the impact on life expectancy is relatively larger than in the general public.

Use of the results

The results can be used to weigh decision criteria in an MCDA. Applied to therapeutic and societal need, the MCDA will give rise to a rank order of diseases in which the diseases with the highest need for a better treatment get the highest score. Before the MCDA can actually be applied, more research is needed on how to score diseases and interventions on each of the criteria included in the model. This work is envisaged for 2015.

CONCLUSION

Disease severity in terms of quality of life under current treatment, and opportunities for improving quality of life through health care interventions are considered to be the most important criteria for resource allocation decisions in health care by the Belgian general population. Compared to the decision makers, the general public attaches relatively less importance to changes in life expectancy.



■ SUMMARY

1 INTRODUCTION

Back in 1977, Slovic et al.¹ stated that “humans are quite bad at making complex, unaided decisions”. Reimbursement decisions in health care are complex, as often characterized by numerous criteria that play a role and should be taken into account when making decisions. Policy makers in a democratic system are moreover faced with the expectation that their decisions are legitimate. As good housefathers, they should make decision in the best interest of the population. The interest can be assumed to be at least partially determined by the preferences of the public. But policy makers most often only have a rough idea about the preferences of the general public, because they do not have data to rely upon. As a consequence they will tend to use intuitive or heuristic approaches to simplify complex decisions, and rely on their own sense of what is best for the population to take decisions.

This study aims to make explicit the preferences of the general public about the relative importance of several decision criteria for the reimbursement of new health interventions. It is yet another step in an extensive piece of research that aims at developing a tool that can serve as an aid in the decision making process and can make the multiple considerations for these decisions more transparent. Rather than to serve as a mechanistic magic formula, the tool should help decision makers making the trade-offs they inevitably have to make between the advantages and disadvantages of reimbursing a new intervention. It should help decision makers to make these trade-offs informed by the preferences of the general public. A first step in the research was taken by KCE in 2010, with the development of a transparent decision framework.² The framework consisted of five relevant questions for every reimbursement decision. It basically splits the decision problem into components that go with a set of coherent criteria. For example, all criteria related to the severity of the disease from the patients' perspective -given the existing treatments for the disease - are clustered into one question, all criteria related to the added value of a new treatment are clustered in another question. This reduced the complexity of the process; considering a limited number of criteria per question is easier than considering all criteria relevant for all questions at once. The questions this study focusses on are:



- Is there a need for a better treatment than the existing (reimbursed) treatment in the targeted condition?
- Is there a societal need for a better treatment than the existing (reimbursed) treatment in the targeted condition?
- Are we as society prepared to pay for the treatment under consideration?

An approach frequently proposed for answering such multi-criteria questions is that of multi-criteria decision analysis (MCDA). MCDA requires a definition of the objective of the decision making process, the identification of the relevant criteria for a decision, the determination of the relative weight of each of the criteria (representing the relative importance of each criterion), the scoring of options on each of the criteria, and the aggregation into an overall score allowing the ranking of different options. The application of the MCDA principles to each of the questions in the decision framework will lead to three weighted scores. For reimbursement to be of high priority, all three weighted scores need to be high, or, if one of the scores is not high, there should be explicit reasons for still granting reimbursement. The question of whether society is prepared to pay *more* for the treatment under consideration, depends on the level of need and the level of added value. *How much* more society is willing to pay is a question that is not included in the current study because the design did not allow for tackling this question in a valid and feasible manner.

The study focusses on the problem of deciding on the reimbursement of *new* health interventions in a situation where the use of public resources needs to be justified and justifiable towards the general public and taxpayers. The existing situation, with current reimbursements, is taken as *is* and not questioned. The current study presents the results of the survey performed to derive public preference weights for a number of reimbursement decision criteria. It also explores briefly how these weights could be applied in a decision tool. Future studies will aim at developing a practical tool for helping decision making, applicable to a variety of conditions, using weights derived from the general public. This requires the development of practical rules for scoring rules options on each of the criteria, based on the available evidence, as well as exploring different possible procedures for applying MCDA.

2 METHODS

2.1 Population survey

A large representative sample of the general public (N=20 000) between 20 and 89 years of age, stratified by age and sex, was contacted by regular mail to participate in a web- or paper survey in February 2014. Of these, 4810 started completing the survey and 4485 (22.4%) answered all choice sets. After two checks of consistency and comprehension, 197 respondents were excluded for analysis and a net sample of 4288 respondents (21.4%) was obtained.

The survey was anonymous. Three reminders were sent to non-responders between February and April, with intervals of 2 weeks.

The survey development process included a pre-test, a pilot test and a test-retest phase. Both the pre-testing and pilot testing were meant to modify the questionnaire in such a way as to improve the comprehension, presentation, and feasibility of the questions. The test-retest, performed in 42 people, showed good overall reliability (Cohen's Kappa = 0.7, approx. 95% CI: 0.62–0.77). Over all choice sets, the majority of the respondents chose the same alternative in test and retest, although the correspondence varies across questions.

The final survey consisted of 9 discrete choice questions, one moral reasoning exercise and a number demographic questions. The body of the survey was structured in three blocks: one relating to criteria that determine the need for a better intervention than the one already available from the patients' point of view (therapeutic need), one relating to criteria that determine the need for a better intervention than the one already available from the society's point of view and one relating to criteria that determine the therapeutic or societal added value of a health intervention. The criteria included in each block have been determined through literature review and expert workshops. With the objective of developing a generic MCDA in mind, the criteria were defined in generic terms. The criteria included in each block are presented in Table 1.

**Table 1 – Criteria included in the survey**

Therapeutic need	Societal need	Added value of new treatment
<ul style="list-style-type: none">• Quality of life with current treatment• Life expectancy with current treatment• Discomfort of current treatment	<ul style="list-style-type: none">• Societal cost of disease per patient• Prevalence of disease	<ul style="list-style-type: none">• Impact on quality of life• Impact on life expectancy• Impact on discomfort of treatment• Impact on disease-related public expenditures per patient• Impact on the prevalence of disease

In each discrete choice question, respondents were asked to choose between two different patient groups (therapeutic need), two different diseases (societal need) or two different health interventions for the same disease (added value). With 24 different versions of the questionnaire, differing in the description of the scenarios between which to choose, and 3 choice sets for therapeutic need, 1 for societal need and 4 for added value, it was possible to obtain weights for each criterion included in a specific block. In discrete choice experiments, which is the technique used for obtaining the preference values from the public, criteria are referred to as attributes, with in each attribute a number of levels. For example, the attribute ‘impact of a treatment on life expectancy’ in “added value” has two levels: (compared to the current treatment), the new treatment does not change life expectancy, or the new treatment increases the life expectancy of patients.

The questionnaire contained a dominant choice set to check the credibility of the responses of the participant. A dominant choice set is a choice set where one of the alternatives presented is superior on all attributes. People should logically choose the dominant alternative. The responses of people not passing this validity check were excluded from the analysis.

The model used to obtain the weights was a main effects multinomial logit model. This model gives coefficient for each level of each attribute. For the multi-criteria decision tool, however, we do not want to be limited to the

levels included in our survey but we need attribute-specific weights that are level-independent. As there is no golden standard for deriving such attribute-specific but level-independent weights, we used two different methods: the log-likelihood method and the coefficient range method in order to test the robustness of the derived weights.

2.2 Decision makers survey

In addition to the population survey, we performed the same survey in public decision making or advisory bodies in Belgium. All members of nine different commissions or councils were invited to participate by e-mail. These respondents were asked to respond as representatives of the group they represent in the committee of which they are a member. A total of 421 representatives received an invitation, of which 175 (41.6%) participated in the survey.



3 RESULTS

3.1 Response and sample characteristics

Of the 4485 respondents who participated in the survey, 4288 (21.4% of the initial sample) completed all choice questions. Their responses were included in the analysis. The distribution amongst sex categories in the sample is similar to that in the general population. The proportion of women amongst the respondents was 52.1%. In the general population, this is 51.3% in the age group of 20 to 89 year olds. Females >70 years of age were underrepresented in the sample, and males and females between 51 and 70 years of age were overrepresented. The large majority of the respondents who answered all choice questions participated through the web (slightly over 91%), although a non-negligible number of respondents cared about asking a paper version (almost 400 people). In view of the finding from the Federal Public Service Economy that 82% of the Belgian citizens regularly access the internet, the bias from having 91% respondents through the internet is expected to be limited.³

The reminders had a marked effect on the response rate: 27% of the responses were received after the initial invitation, 34% after the first reminder, 19% after the second reminder and 20% after the third reminder. In the age groups 70-79 and 80-89, the proportion of people requiring 3 reminders before participating is typically higher than in the other age groups. This also applies to the lower educational groups (from “did not finish primary school” to “finished lower secondary school”), which required more reminders than the higher educational groups (from “finished upper secondary school” to “finished university”).

In the group of decision makers, 175 (41.6%) responded to the survey and 161 (38.2%) answered all choice sets. All completed the survey in the web interface. About 57% of the respondents chose the Dutch survey and about 43% chose the French survey. All the advisory committees of the RIZIV / INAMI had a participation rate of >45%. The parliamentary committee for health and the senate committee on social affairs had the lowest response rate (11.6% and 9.3% respectively).

Eleven percent of the respondents in the general population sample reported having a serious illness; 32.3% reported to have a relative with a serious illness. In the decision makers' sample, this was 5% and 39.4%

respectively. None of the decision makers rated his/her health as bad or very bad. In the general population sample, a small minority rated his health as bad (4.1%) or very bad (0.6%).

3.2 Consistency and certainty of responses

Less than 1% of the respondents systematically chose always the first alternative or systematically chose always the second alternative in the nine choice sets. This is the case for both the general population and decision maker sample. About 96% of the general population sample and over 99% of the decision maker sample chose the ‘dominant’ alternative in the dominant choice set introduced in the added value block as a credibility check. Respondents who did not choose the dominant alternative were excluded from the analyses.

Respondents were in general quite certain about their choices. For 72% (therapeutic need) to 76% (added value) of the responses, people indicated they were either very certain or certain about their response.

3.3 Modelling results

3.3.1 Therapeutic need

Both the public and the decision makers gave the highest weight to quality of life with current treatment. Both groups consider the therapeutic need to be the lowest in people with a good quality of life given current treatment, who do not die from their disease and with little treatment discomfort. Therapeutic need is considered to be the highest in patients who die from their disease, experience much discomfort from current treatment and have a low quality of life.

A few additional observations can be made:

- People do not seem to make a difference between “dying 5 years earlier than patients without the disease” and “dying immediately from the disease”. Both features of a disease have a similar impact on the valuation of the therapeutic need. This means that people, when confronted with such difficult choices, tend to dichotomize between “lethal” and “non-lethal” diseases.



- When all else is equal, respondents basically choose on the basis of age of the patient. For example, when two patient groups are in a different lethal health state, both have a low quality of life and little discomfort of current treatment, people express a preference for the group that happens to be younger on average. This could be explained by the fact that younger patients would lose more life years when confronted with a lethal disease than older patients, and by the fact that duration of the health benefits (in terms of quality of life or life expectancy) of a new intervention is expected to be longer in younger patients because they are expected to have more years ahead of them than older patients. Nevertheless, the distinction between <18 year old patients and patients between 18 and 64 years of age is not explicitly brought to the fore in the responses of the survey participants. Compared to patients between 18-64 years of age, the therapeutic need in patients aged <18 years is only a little bit higher, if all other disease and current treatment characteristics are kept equal.
 - Respondents valued avoiding quality of life losses in patients that are currently in rather good health higher than avoiding the same absolute quality of life loss in patients who already have a low quality of life.
 - Citizens seem to make trade-offs between life expectancy with current treatment and quality of life when judging therapeutic need. All else equal, people consider the therapeutic need in a disease with a quality of life of 5/10 from which patients die 5 years earlier than people without the disease as equal to the therapeutic need in a disease with a quality of life of 2/10 from which patients do not die prematurely. In other words: a lower quality of life and a low impact on life expectancy is equivalent to a better quality of life and a higher impact on life expectancy.
- The results of the model for the decision makers are different from those of the general public in two important ways:
- Decision makers have a higher preference for developing a new intervention for patients between 18-64 years old than for patients younger than 18 years of age, if all else is equal.
 - Decision makers do make a distinction between “dying 5 years earlier from a disease” and “dying almost immediately”. Patients suffering from a disease causing immediate death are considered to have a higher therapeutic need than patients suffering from a disease that decreases life expectancy with 5 years. However, the difference in the level of therapeutic need between a lethal disease and a non-lethal disease is higher than between a disease from which patients die 5 years earlier and a disease from which patients die almost immediately, meaning that decision makers also make a clear distinction between lethal diseases and non-lethal diseases in judging therapeutic need.

Comparisons between the general public and the decision makers should be treated with caution, though, as the sample of decision makers is rather small and therefore the observations for this group more uncertain (i.e. confidence intervals around the estimated coefficients and weights are larger).

The implicit weights given to the criteria included in the therapeutic needs domain by the general public sample and by the decision makers sample are presented in Table 2. The results of the two methods applied for deriving weights are presented separately. The weight should be interpreted as reflections of how important each criterion should be in the appraisal of the therapeutic need in a patient population with a particular disease.

The relative importance of the three attributes for therapeutic need differ between the decision makers and the general population, not only in relative weight but even in the order of importance. The general public finds discomfort of current treatment more important than impact on life expectancy, while decision makers find life expectancy more important than discomfort of current treatment.

**Table 2 – Weights for criteria in the Therapeutic Need domain**

	General population		Decision makers	
	Log-likelihood method (rang)	Coefficient range method (rang)	Log-likelihood method (rang)	Coefficient range method (rang)
Life expectancy	0.14 (3)	0.22 (3)	0.32 (2)	0.34 (2)
Quality of life	0.43 (1)	0.42 (1)	0.53 (1)	0.45 (1)
Discomfort	0.43 (1)	0.36 (2)	0.15 (3)	0.21 (3)

3.3.2 Societal need

The results of the model for societal need show different results depending on the method used to derive weights. With the loglikelihood method, the rank order of the criteria for societal need differ between decision makers and the general public. With the coefficient method, however, the rank order

is the same. With the latter method, it is found that the general public and the decision makers attach more importance to the prevalence of a disease than to the impact of the disease on public expenditures per patient when assessing the societal need for a better treatment (Table 3). Both groups consider the need to be highest in very frequent diseases that cost a lot to society per patient.

Table 3 – Weights for criteria in the Societal need domain

	General population		Decision makers	
	Log-likelihood method (rang)	Coefficient range method (rang)	Log-likelihood method (rang)	Coefficient range method (rang)
Public expenditure	0.65 (1)	0.45 (2)	0.44 (2)	0.34 (2)
Prevalence	0.35 (2)	0.55 (1)	0.56 (1)	0.66 (1)



3.3.3 Added value of new treatments

New treatments are considered to have an added value if they reduce public expenditures, improve quality of life, increase life expectancy, reduce treatment discomfort or reduce the prevalence of the disease. The added value is considered to be influenced most by changes in quality of life.

A general observation is that the value loss associated with something negative (higher expenditures, higher treatment discomfort, less patients cured) is higher than the value gain associated with something positive (lower expenditures, lower treatment discomfort, more patients cured). For example, the negative effect on the perceived added value of increasing public expenditures is higher (-0.43) than the positive impact of decreasing public expenditures (+0.23). This means that people's preference *against* interventions that increase public expenditures is stronger than their

preference *for* interventions that decrease public expenditures. Otherwise stated: there is less to be gained in terms of added value from choosing a cost-saving intervention than from avoiding a cost increasing intervention, according to the general public's point of view. The same applies to impact on treatment discomfort and on prevalence of a disease.

For quality of life, this is less clear: the gain in added value associated with increasing quality of life is about the same as the loss associated with reducing quality of life, disregarding the current quality of life of patients.

For the general public, reductions in the prevalence of a disease play an almost equally important role in the assessment of the added value as changes in quality of life. Decision makers also consider reductions in prevalence to be of second most importance, but not equally important as changes in quality of life.

Table 4 – Weights for criteria determining the added value of new treatments

	General population		Decision makers	
	Log-likelihood method (rang)	Coefficient range method (rang)	Log-likelihood method (rang)	Coefficient range method (rang)
Change in quality of life	0.37 (1)	0.30 (2)	0.39 (1)	0.32 (1)
Change in prevalence	0.36 (2)	0.31 (1)	0.29 (2)	0.28 (2)
Change in life expectancy	0.14 (3)	0.15 (3)	0.21 (3)	0.19 (3)
Impact on public expenditures	0.07 (4)	0.12 (4)	0.08 (4)	0.13 (4)
Impact on treatment discomfort	0.06 (5)	0.12 (5)	0.03 (5)	0.07 (5)



3.4 Weights by population subgroups

For the ultimate purpose for which the weights were measured, it is less important what the differences in preferences are between population subgroups. Nevertheless, sub-group analyses were performed to examine whether preferences differed between population subgroups and to assess the risk of bias in our results if our survey sample would not be representative. It has not been possible for all variables of interest to check whether our survey is representative, because there are no national data available but only proxies with their own limitations.

When sub-groups defined by age category of the respondent are compared, we observe that the 80-89 year olds clearly have different preferences than the other age groups.

For therapeutic need, respondents between 80 and 89 years of age give much more importance to the criterion of discomfort of current treatment and less to the criterion of quality of life under current treatment than the other age groups. For societal need, the 80-89 year olds give a higher weight to prevalence than to public expenditures in judging societal need, unlike all other age groups give a higher weight to public expenditures. As for the judgment of the added value of new treatments, the 80 to 89 year olds give relatively more weight to improvements in quality of life than the other age groups. At the same time, changes in treatment comfort are more important than changes in life expectancy for this group as well as for the 70-79 years old. This means that these age groups value living better more than living longer, whether "better life" is defined by better quality of life or less treatment discomfort. In contrast, the other age groups typically give more weight to improvements in life expectancy than to reductions in discomfort, but they also give more weight to improvements in quality of life than to increases in life expectancy. The respondents in the youngest age group (20-29y) give relatively more weight to reductions in public expenditures compared to the other age groups, although this criterion also for this age group remains the least important for the assessment of the added value of a new intervention.

People who report being currently in good health give slightly more weight to quality of life when judging therapeutic need than to discomfort of current treatment. Respondents who report not being in good health find it more important to reduce treatment discomfort than to increase overall quality of life. Both subgroups give the lowest weight to reductions in life expectancy due to the disease.

3.5 Using the weights

The measurement of the relative importance of different criteria in the decision making process is not the ultimate endpoint of our research endeavour. The weights are just one input into the multi-criteria decision analysis (MCDA) tool we are developing to support health care reimbursement decision making.

MCDA involves (1) scoring diseases and treatments on a number of selected criteria, and (2) weighting these scores with the weights reflecting the relative importance of each of the criteria, and (3) summing the weighted scores to obtain an overall score reflecting the level of need for a new treatment and the level of added value of a particular new intervention. The scoring is done by the members of the appraisal committees, after careful consideration of the available scientific evidence with respect to each of the criteria relative to the disease and treatment under consideration. The weights obtained from this study could be used for the weighting of each of the scores. The weights remain constant across decisions, i.e. the same weights are applied to the criteria, independent of the disease or treatment under consideration, only the scores will differ case by case as they are disease- and intervention-specific. The weights simply indicate to what extent a criterion should be taken into account in the decision making process. They are, as such, not reflecting the clinical significance of a particular level of a criterion. The clinical significance is reflected in the scoring, the weights indicate to what extent a clinical significant or insignificant effect should matter for the decision.

The weights of the general public for each of the criteria in each domain, as derived using different methods, are summarized in Table 5.

**Table 5 – Weights of the criteria in each cluster**

Cluster	Decision Framework Question	Criterion	Weights*
Therapeutic need	Is there a need for a better intervention for this condition than the best intervention currently reimbursed from the patients' point of view?	• Life expectancy, given current treatment	0.14 – 0.22
		• Quality of life, given current treatment	0.43 – 0.42
		• Discomfort of current treatment	0.43 – 0.36
Societal need	Is there a need for a better intervention for this condition than the best intervention currently reimbursed, from the society's point of view?	• Prevalence of the disease, given current treatment	0.35 – 0.55
		• Public expenditures associated with the disease, per patient, given current treatment	0.65 – 0.45
Added value	Are we, as society, prepared to pay for this particular intervention out of public resources?	• Impact of the new intervention on life expectancy, as compared to current treatment	0.14 – 0.15
		• Impact of the new intervention on quality of life, compared to current treatment	0.37 – 0.30
		• Impact of the new intervention on the discomfort of treatment compared to current treatment	0.06 – 0.12
		• Impact of the new intervention on prevalence of the disease, as compared to current treatment	0.36 – 0.31
		• Impact of the new intervention on the disease-related public expenditures, as compared to current treatment	0.07 – 0.12

* The first figure is the weight as derived with the log-likelihood method, the second figure is the weight derived with the coefficient range method.



3.6 How to apply the multi-criteria decision tool

The target users of the MCDA tool are the committees within the RIZIV – INAMI who have to give advice to the Minister about the reimbursement of new products or services. The basic idea is that committee members apply a MCDA on each of the three questions mentioned in Table 5 whenever reimbursement is requested for a new treatment. In this application, they should use the criteria corresponding with each question with their respective weights.

The MCDA is applied as follows:

- **Step 1: Consideration of the condition targeted by the new treatment and the current treatment for the condition.**

The committee members consider the condition targeted by the new treatment and score the criteria relating to therapeutic need (quality of life under current treatment, current treatment discomfort and impact of the disease on life expectancy despite current treatment) and relating to societal need (prevalence of the condition, and average public expenditure per patient with the condition).

For the scoring, the committee members should dispose of an assessment report describing the existing scientific evidence regarding each criterion, as well as the evidence gaps. The members could consult external experts, e.g. in case of insufficient or inconclusive evidence. Scoring rules and procedures will be developed in a future KCE report.

- **Step 2: Consideration of the added value of the new intervention**

The committee members score the criteria for added value for the new intervention for which reimbursement is being considered; they score the impact of the new intervention on the quality of life of patients, on treatment discomfort, on life expectancy, on public expenditures per patient and on the prevalence of the condition, each time compared to the current situation with the currently available treatment. As for therapeutic and societal need, the scores should be based on the best available scientific evidence.

- **Step 3: Weighting of scores for therapeutic need, societal need and added value**

The scores given to each of the criteria included in the MCDA model are weighted with their respective public preference weights, as derived from the current study (last column of Table 5). This is done by multiplying the score with the weight. As different methods revealed different weights and there is no theoretical basis to assume that one method is better than the other, it is important to choose one single set of weights and use the same set across different appraisals. This emphasizes once again the importance of not using MCDA as a single magic formula that provides an easy solution to complex problems.

For each domain (therapeutic need, societal need and added value) the weighted scores of the domain-specific criteria are summed. This results in three scores: one for therapeutic need, one for societal need and one for added value of new treatment. Higher scores represent a higher level of priority in terms of therapeutic need, societal need or added value of treatment, depending on the domain considered. By repeating the MCDA for different decisions, a priority ranking of diseases and treatments will eventually be obtained.

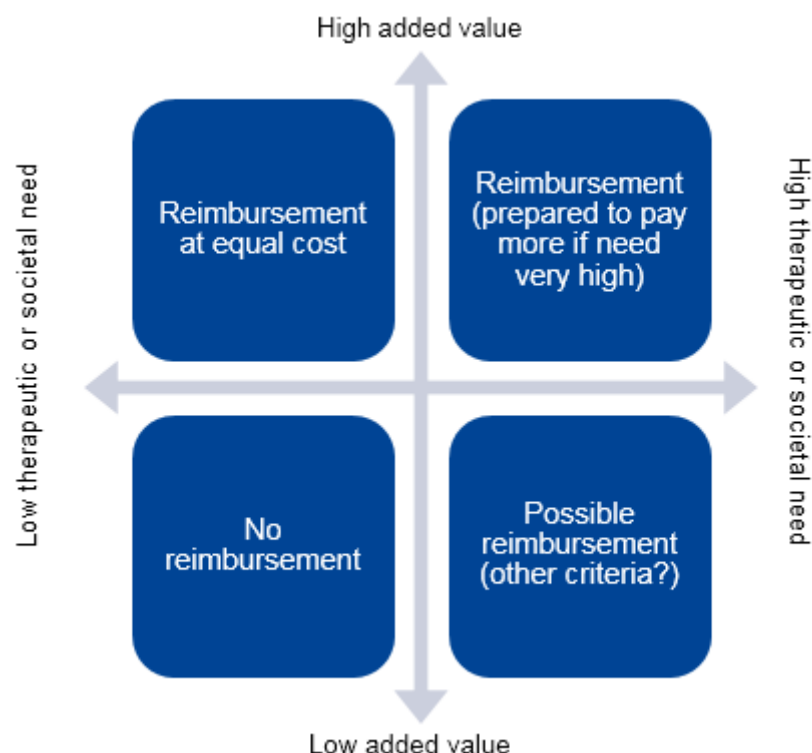
- **Step 4: Deliberation about the resulting scores for therapeutic need, societal need and added value.**

Once the three sums of weighted scores have been calculated, the commission has to consider in which quadrant of Figure 69 the intervention is located. The higher the need and the higher the added value, the more likely it is that reimbursement can be considered. This is a relatively simple decision rule. However, there might be criteria that are not included in the MCDA that matter to the decision. The deliberation should therefore include a discussion about whether there are other criteria -not yet included in the MCDA- that are important and that would justify a change in the priority ranking in terms of need or in terms of added value. For example, it could be that policy makers wish to give higher priority to prevention than to cure. If that is the case, preventive interventions might be moved up in the ranking in order to get higher priority. If other criteria are considered important, they should be made explicit and the committees should explain how these additional criteria modified the ranking of a disease or a treatment.



Eventually, also the budget impact of a treatment will determine whether reimbursement is acceptable or not in a given budgetary context.

Figure 1 – Preparedness to pay (more) for a new intervention



4 DISCUSSION

The MCDA tool is an aid to decision making, not a guideline. Crucial for any policy maker, making decisions on behalf of and for the benefit of society, are the ethical considerations. Policy makers are elected for applying general ethical principles when making decisions. The ethical reflections are to be made independent of any particular decision, as they are supposed to apply generically, meaning that they should be defined a priori, i.e. before the execution of the MCDA. That is why the process should not stop at the point where the three scores have been calculated but should be complemented with a deliberation in which potential additional considerations (clearly distinct from the criteria already included in the MCDA!) should be discussed. One of the benefits of the MCDA is that it makes the transparent how the criteria are taken into account in the decision making process. This transparency requirement should also apply to the additional considerations that are used to change the outcome of the MCDA. If not, the decision process will remain opaque and it will be unclear whether the preferences of the population eventually really mattered.

The MCDA application developed in this study is different from many examples of MCDA models in literature, in that we propose to apply an MCDA to each cluster of criteria, being therapeutic need (disease-related criteria from patients' point of view), societal need (disease-related criteria from the societal point of view) and added value (intervention-related criteria). Many MCDA models described in literature aim at one weighted score covering all relevant clusters. We have several reasons to suggest the use of a multi-layered MCDA.

First, we presumed that the willingness to reimburse a new treatment out of public resources would probably be a function of the level of therapeutic and societal need. A new treatment with a presumably high added value, could still be not worthwhile to reimburse because there simply is no need for a new treatment. Our model therefore foresees the calculation of three weighted scores: one for therapeutic need, one for societal need and one for added value. In case of a high need and a high added value, decision makers will be more inclined to consider reimbursement than in case of a low need and a low added value. However, there are several situations in which a conditional decision might be taken. For example, in case of a low therapeutic and societal need and a high added value, the authorities might



still want to reimburse a new intervention, under the condition that the overall cost of the treatment is the same as that of the comparator. An economic evaluation can provide this information. Or, when no active alternative treatment is available but the only available alternative is best supportive care, decision makers might still want to reimburse a new promising intervention, with currently a limited added value, to keep the door open for further improvements in the development of the intervention. In such cases, often specific conditions for reimbursement will have to be defined (i.e. who gets reimbursement, under which conditions) and a re-assessment after some time will have to be scheduled.

Second, by creating a stepwise hierarchical decision-making process, the number of criteria per step in the process diminishes as compared to an all-encompassing one-step decision-making process. It makes the considerations more manageable from the cognitive point of view.

An important next step is the development of scoring rules for the criteria in the MCDA tool. These will have to include ways to deal with missing or low-quality evidence. Scoring rules should be followed by all committees that use the tool, to ensure consistency. The more the MCDA is used in decisions, the more useful will become, because then it will become possible to refer to previous applications of the tool when considering the level of need for a better treatment and the level of added value of a new treatment.

Pilot studies, testing MCDA for reimbursement decisions in other countries, found that decision makers generally perceive the technique to be useful as a decision support. In particular the systematic consideration of multiple decision criteria in a pragmatic way is felt to be useful and to improve the decisions. VTS-HTA, the HTA agency in Lombardy (Italy) implementing reimbursement decisions (mainly medical devices and diagnostics), is currently systematically using an MCDA framework to decide on reimbursement.

Our study has demonstrated that more research is needed on methods to derive criteria weights. Different methods, of which we have applied two in the current study, give different weights. This gap has already been highlighted by other researchers. Because MCDA is not an exact science, the point value of the weights is less important than their relative importance or rank amongst the full set of criteria. But if different methods give different rank orders of criteria, more research is needed to find out which method gives the best results in terms of acceptance of the disease and intervention rankings resulting from their application.

5 CONCLUSION

MCDA is not a formula that leads to “easy” yes/no decisions. It is only through the consistent use and consideration of the relevant questions with the relevant criteria and their relative weights, that the decision-making process can become more consistent. Consistency implies rationality, in the sense that decisions about what the budget allows are more in line with what people consider important, both for individual patients as for the society as a whole. Moreover, the MCDA allows more transparency in the process. The remit of the advisory committees remains the same. The committee members remain responsible for the appraisal of the new interventions on different criteria. The only difference will now be that the weights given to each of these criteria will be those of the general public and not those of the committee members. Committees will still, as before, discuss additional criteria that are not included in the MCDA framework presented in this study and will still have to formulate an advice based on their appraisal. However, we hope this framework can help to justify advices towards to general public and to create, as such, a societal ground for the decisions made.



■ SCIENTIFIC REPORT

1 BACKGROUND AND SCOPE

1.1 Scope of the study

Several issues arise when making decisions about the reimbursement of new health care interventions (procedures, services, drugs). First, a large array of criteria of various nature, often fraught with mutual contradictions, are relevant and have to be weighed and considered. Reimbursement decision-making not only requires “technical” judgments, such as those on safety, clinical effectiveness and organisational issues, but also involves “value” judgments,⁴ requiring the weighting of the technical judgments. This weighting of decision criteria is seldom straightforward.

Second, the reimbursement of one particular intervention is in competition with any other intervention or any other use of the resources needed to reimburse that intervention if the budget is limited. Taking a broader societal stance, considering the trade-offs between health care interventions but also between health care and other social services, supported by societal needs and preferences, is difficult.

Finally, in a rapidly evolving world, today’s decision logic is not necessary applicable tomorrow.

The primary focus of this report is on decisions about the **reimbursement of new products or services for a particular patient population**, which are preceded by a structured assessment of the evidence regarding safety, efficacy, effectiveness, costs, patient-related issues and applicability. This is typically the case for drugs and medical devices, but could also be applied to other specialised services (e.g. dental care, mental health care, screening services, etc.), as long as there is a clear definition of a disease, an intervention and the alternative for the intervention (current care). In Belgium, the Drug Reimbursement Committee (CTG – CRM) and the Reimbursement Committee for Implants and Invasive Medical Devices (CTIIMH – CRIDMI) give advice to the minister regarding the reimbursement of drugs and medical devices, respectively. For health care services, there are also advisory or decision-making organs within the National Institute for Health and Disability Insurance (RIZIV – INAMI), for instance for revalidation and for dental care. In theory, the relevance of the product created in the current study –being a tool to support reimbursement decisions- is broader. The tool could in principle also be applied, for instance, to disease



management strategies or organisational changes in health care, but this would require further reflection as the decision criteria might in these cases be different. Currently, we focus on decisions about the reimbursement of new products or services for a clearly identifiable health condition for which the current treatment can also be defined.

Reimbursement decisions for a particular intervention are usually based on an individual dossier – i.e. a dossier regarding that particular intervention without the broader context of health care in general - and are taken on a case-by-case basis. They are essentially incremental in nature, meaning that they consider and weight the incremental value of an intervention as compared to the standard treatment currently available. Several criteria can be considered relevant during the reimbursement decision-making process: disease severity, therapeutic need, personal responsibility for the disease, treatment safety, clinical effectiveness, cost-effectiveness and budget impact.

The appropriateness of the current reimbursement of products and services is not questioned in this study. The aim is to provide a tool to help policy makers in deciding which **new** interventions deserve high, medium or low priority for reimbursement. The presumption is that a reimbursement decision has to be taken (usually within specific time limits) whenever a request for reimbursement is submitted. Moreover, we presume that the objective of the decision maker is to take decisions that are supported by the general population.

Key points

- **The scope of this study is limited to decisions about the reimbursement of *new* health care products or services, which are targeted at a clearly identifiable and specific health condition, and for which the current treatment or condition management alternative can be defined.**
- **The appropriateness of current reimbursements is not questioned.**
- **It is assumed that a decision has to be taken when a reimbursement request is submitted to the RIZIV – INAMI and that policymakers aim at taking decisions supported by the general population.**

1.2 Legitimate decision-making

1.2.1 *Accountability for reasonableness*

According to the ethical-theoretical framework for accountability for reasonableness developed by Daniels and Sabin, a legitimate decision process requires transparency in the criteria used for formulating a reimbursement advice or making a reimbursement decision, relevance of decision criteria, revisability of decisions and enforcement of the transparency, relevance and revisability conditions.⁵

In a social health insurance context, where health care is funded mainly from public resources, relevance of the decision criteria means that the criteria used for making decisions are supported by the society, i.e. are considered relevant from a societal point of view. It has been argued that “any policy that strays too far from what is acceptable to a broad spectrum of health care consumers and providers and the general public will not succeed.”⁶ This does not imply, however, that the majority’s view should always be followed. For example, if the majority’s view is unethical or unconstitutional, policymakers may decide not to follow it.

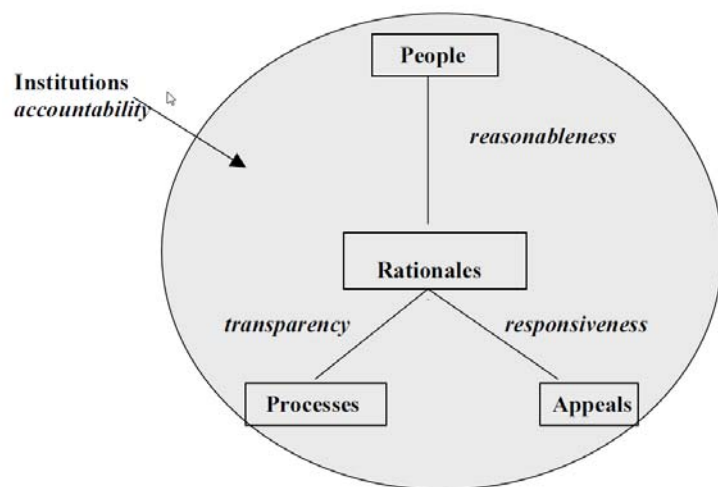
Transparent application of these criteria would in this context mean that the relative weight given to each of the criteria is made explicit and that the process used to establish these weights is clear. Relevance would mean that the applied weights match their relative importance according to the society. For example, if having a good life is considered more important than having a long life, then this should be reflected in the decisions taken (e.g. palliative treatments should get higher priority than treatments that extend life in a very bad health state).

Revisability means that decision makers should revise reimbursement decisions when new evidence becomes available or preferences or values with respect to health interventions change.

1.2.2 Putting accountability for reasonableness into practice

A useful model to put the concept of accountability for reasonableness with its conditions into practice has been developed by Gibson et al. (2002)⁷ The model is presented in Figure 2.

Figure 2 – Model of accountability for priority setting in health care



Source: Gibson, 2002⁷

The model has been developed as a response to the observation that decision makers have little guidance to help them make their priority-setting decisions. On the one hand, discipline-specific ethical approaches to priority setting take insufficiently the actual experiences of decision makers faced with conflicting ethical values for priority setting into account. On the other hand, empirical descriptions of how priorities are set in real-life are insufficient because they tell nothing about how priorities *should* be set. The authors state that *“reasonable people, having diverse moral views, disagree about what constitutes a fair allocation of resources to meet competing health care needs. In this absence of consensus on guiding principles, the*

*problem of priority setting becomes one of procedural justice – legitimate institutions using fair processes.”*⁷

It is the procedural justice at which the accountability for reasonableness framework of Daniels and Sabin aims.⁵ The model integrates both the empirical realities of how decisions *are* made (e.g. of institutions making decisions on behalf of the society) with the ethical values of how decisions *should* be made.⁷ At the centre of the model is “rationales”, encompassing both the factors determining a decision (e.g. safety, effectiveness, cost-effectiveness) and reasons for a decision (taking the different relevant factors and the relevance of these factors into account). If people (stakeholders, experts) contribute to the rationales, reasonableness will be improved. If the decision-making processes allow decisions and rationales to be compared with previous decisions and rationales to ensure consistency, the rationales will be more transparent. And finally, the possibility to appeal against a decision will increase the responsiveness of the system. The institutions are the bodies within which everything happens. By means of people involvement in the decision-making process, clear and transparent processes and rationales and appeal mechanisms, the institutions can reach accountability for reasonableness.

The current study focuses on the “people” and “reasonableness”-part of this model, i.e. trying to identify possible rationales for decisions based on people’s values and preferences. Reasonableness is considered to be an operational goal, as are transparency and responsiveness.⁷

1.2.3 Accountability for reasonableness in a deliberative decision-making system

In 2010, the KCE published an international comparison of drug reimbursement systems. The report contained a set of recommendations towards policymakers to improve the accountability for reasonableness of drug reimbursement systems by improving the transparency and relevance of drug reimbursement processes.² A reimbursement decision-making process essentially consists of three phases: the assessment, the appraisal and the decision. In the assessment phase the evidence regarding the technology under consideration for reimbursement is collected. In the appraisal phase, this evidence is considered and weighted. Appraisal implies value judgments, e.g. related to the relative importance of each of the assessment elements. These value judgments should, in a democratic



system, ideally reflect societal values and preferences. In the decision phase, a decision is made based on the outcome of the appraisal.

It could be argued that a deliberation-driven system, as in Belgium, where all stakeholders are represented in the appraisal committee, leads to decisions that are consistent with public preferences. However, there are very little opportunities to prove this if the criteria are not defined explicitly and their relative importance is not defined. Devlin and Sussex (2011, p12) state that *“Deliberative processes carry a risk of unintended inconsistency in the way qualitative judgments are made across conditions, technologies and patients”*.⁸ They cite Baltussen and Niessen (2006, 2-3) who wrote: *“When confronted with such complex problems, policy makers tend to use intuitive or heuristic approaches to simplify complexity, and in the process, important information may be lost and priority setting is ad hoc ... policy makers are not always well placed to make informed, well-thought choices involving trade-offs of societal values.”*⁹ As a consequence, the metaphor of the black box applies to both the decision makers and the public observing this process.

1.2.4 Preferences and values

It is important to make a distinction between preferences and values. Preferences relate to individuals' wishes, whereas values are societal and relate to “what ought to happen”. When deciding based on the preferences of the majority of the society, decision makers might come in conflict with societal values. Societal values have an ethical dimension that may surpass individual preferences. They can relate to, for instance, general ethical principles of non-discrimination or no blame, or the respect for individual's preferences over their own situation. Societal values are guiding principles for decision-making. However, they are often difficult to operationalise and are as such insufficient to give a straight answer about what is the most legitimate decision in one particular case.⁷ When considering the reimbursement of one specific intervention, a decision maker needs to weigh the different advantages and disadvantages of that particular intervention. General ethical values may support and inspire this weighting exercise but are little concrete about how to do this. This is where evidence on public preferences about these features becomes particularly relevant for the decision maker. Information about what the public finds to be more or less important for the reimbursement can help the decision maker, who has to

decide on behalf of the public, to do the weighting exercise. The underlying assumption for this study is that the decision makers, as representatives of the population, have been chosen for the general ethical principles they defend (they represent the values of the political party that has been chosen by the population to take decisions on their behalf). When making decisions about the reimbursement of health care interventions, they apply these general ethical principles, but in addition need more concrete information on what the population prefers when choices have to be made that involve trade-offs between different features of interventions (e.g. would the population rather chose an intervention that prolongs life or improves quality of life, given a particular disease?). This kind of concrete questions cannot be assumed to be covered by the votes people make when elections are held. For these, additional information is needed. The focus of this report is therefore on the general public's preferences for trade-offs between concrete decision criteria rather than on general moral and ethical principles (i.e. values) that should be applied when making reimbursement decisions. We use the term ‘preferences’ to refer to the relative importance of specific criteria and ‘values’ to refer to the more general moral and ethical principles, even though it is acknowledged that also the preferences of the general public with regard to decision criteria are based on their moral and ethical values.

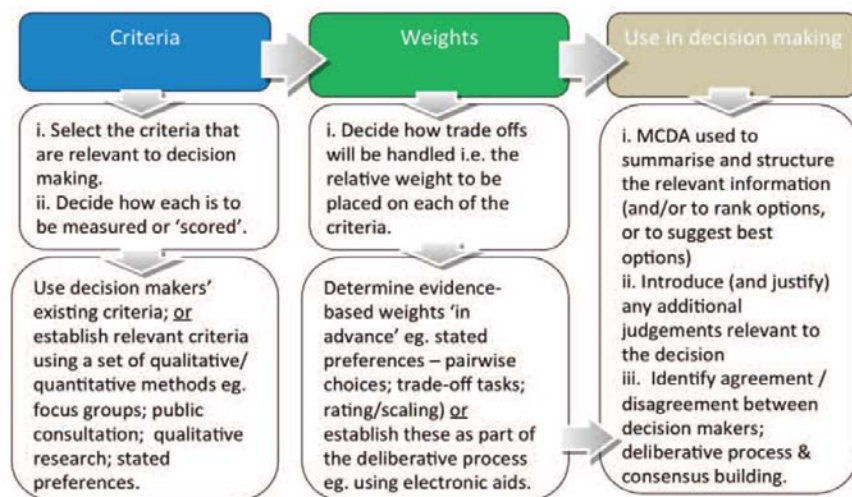
1.3 Multi-criteria decision analysis (MCDA)

1.3.1 Structuring the debate and making decision processes transparent

Preferences for reimbursement of specific health care interventions are determined by multiple criteria or characteristics of patients, interventions, diseases and societal contexts. To allow consistency across decisions, ways to deal with multi-criteria problems are needed. A variety of methods and approaches for taking multiple criteria into account in decision-making have been developed under the heading of “multi-criteria decision analysis” (MCDA). MCDA can be defined as *“a set of methods and approaches to aid decision-making, where decisions are based on more than one criterion, which make explicit the impact on the decision of all the criteria applied and the relative importance attached to them.”*⁸.

Figure 3 presents the key steps for developing an MCDA framework.

Figure 3 – Steps to be taken in the development of an MCDA framework



Source: Devlin et al. (2001)⁸

MCDA should be considered as a *support* to decision-making, helping to structure decisions and deliberative processes that involve multiple criteria. It should not be seen as a prescriptive tool but rather as a tool to exercise the judgments made during multi-criteria decisions. Hence, MCDA is not a mechanistic decision tool. As Klein stated in 1993: “*what really matters is how that debate is structured: how far it promotes reasoned, informed, and open argument, drawing on a variety of perspectives and involving a plurality of interests.*”¹⁰ The issue is not to resolve the dilemma of setting priorities in health care, but rather to structure the decision process and include the preferences of society.

The major advantage of MCDA is increased transparency and consistency in health technology appraisal processes, leading to a decision. The discipline of explicit identification and weighting of the criteria upon which

health care resource allocation decisions are made is valuable. MCDA makes it easier to hold decision makers accountable for the decisions they make on behalf of the general public. Therefore, MCDA should lead to greater public confidence in these decisions.⁸

1.3.2 A five-question framework

The previously mentioned KCE Report 147 presented a possible MCDA framework for making health technology appraisal processes more transparent (Table 6).² The framework consists of five questions that have some kind of hierarchical relationship. If the response to the first question is clearly “yes”, decision makers can go on to the next question. However, more often than a clear “yes” or “no”, the answer will be a gradation of “rather yes” or “rather not”. Therefore, the relationship will not be purely hierarchical. Evaluations on previous levels will be used to develop answers for the next levels. For example, for a new intervention to be worthwhile (acceptable for reimbursement), there should at least be a therapeutic and/or societal need for another intervention in this indication. If people do not feel a need for another intervention, no public resources should be spent on this intervention, unless it is equally good as the intervention that exists but is less costly. However, it is not enough that there is a perceived need. Even if there is a need, the new intervention still needs to be better on criteria that are important. The higher the need and the better the intervention is on criteria that matter to patients and citizens, the higher the propensity to pay for the new treatment out of public resources. Hence, how much society is prepared to pay for a new intervention depends on both the level of therapeutic and societal need for a better intervention and the level of added value of the intervention.



Table 6 – Key questions and possible criteria for a drug reimbursement appraisal process (MCDA framework)

Decision	Question	Possible criteria
Medical, therapeutic and societal need	Does the product target a medical, therapeutic and societal need?	Medical need: Life-threatening / non-life threatening condition; Severe / mild symptoms; Poor initial health state Therapeutic need: Effective alternative treatments available / not available Societal need: High / Low prevalence; Health inequality; Baseline health level
Preparedness to pay out of public resources for a treatment	Are we, as a society, in principle, prepared to pay for a treatment that will improve this indication out of public resources?	Own responsibility Life-style related condition
Preparedness to pay out of public resources for the treatment under consideration	Are we, as a society, prepared to pay for this particular treatment, given that we in general would be prepared to pay for a treatment for this indication?	Safety and efficacy of the treatment compared to the alternative treatment(s); Curative, symptomatic, preventive; Therapeutic value; Significance of health gains
Preparedness to pay more	Given that we, as a society, are prepared to pay for this treatment out of public resources, are we prepared to pay more for this treatment than for the best alternative treatment?	Added therapeutic value; Potentially induced savings elsewhere in the health care sector; Quality and uncertainty of the evidence ; Acceptability of co-payments and/or supplements; Rarity of disease
Willingness to pay (price and reimbursement basis)	How much more are we willing to pay out of public resources for this particular treatment?	Added therapeutic value; Budget impact / ability to pay; Cost-effectiveness ratio; Medical, therapeutic and societal need; Quality and uncertainty of evidence; Limits to cost sharing



Each question needs to be answered using explicit decision criteria. The criteria must be (1) relevant and (2) weighted in accordance with the relative importance attached to them by the general public. The relevance of the criteria and the weights are unknown for Belgium. In literature, several frameworks have been described for other countries with more or less extensive lists of relevant criteria, but the weights are rarely based on public preferences.¹¹ Moreover, the criteria are rarely structured in a decision framework with levels as in the KCE framework. More often, large sets of criteria are introduced at once in an extensive decision model. Our layered model has a limited number of criteria on each level, because the literature shows that it is cognitively almost impossible for people to consider more than six criteria at once when making a choice.¹²

1.3.3 MCDA in practice

Figure 4 shows how the multi-criteria decision model could eventually be used in practice. The procedure starts with reimbursement requests for a new health intervention. The scientific evidence with respect to the new intervention must first be described in a(n) (health technology) assessment report by scientific experts.

Next, the scientific evidence must be appraised by a group of stakeholders and experts (including patients). Appraisal implies weighting the diverse pieces of evidence relating to different criteria described in the assessment report, and making an overall judgment about the desirability of reimbursing a particular product or service. In MCDA, appraisal first involves the scoring of criteria on a predetermined scale, e.g. a Likert scale going from “no impact” to “major impact”. Clear rules with respect to which kind of evidence leads to which kind of score should be established to ensure consistency. Subsequently, these scores need to be weighted in such a way that they reflect their relative importance.

So far, public preferences have not been taken into account in the appraisal process. The next step brings in the preferences of the general population. This happens by weighting the scores assigned to the different outcomes of the intervention with the relative preference for each of these outcomes according to the general public.

The weighted scores can then be summed to obtain an overall score indicating the level of priority. The higher the weighted sum is, the higher is the priority of the intervention and the more important the reimbursement. If

criteria that are not included in the MCDA framework are considered relevant for a particular technology, they first need to be defined and scored. It is important to be explicit about these additional criteria and report them with their scores to maintain the benefit of transparency of the framework. Although we consider the scoring and weighting with public preference weights to be part of the appraisal process, it could also be argued that – given the use of empirical data for both the scoring and the weighting of criteria- this becomes an extension of the assessment process and that the real appraisal starts with the consideration of criteria that are not yet included in the MCDA.

Figure 4 – Multi-criteria decision analysis: how it could work in practice





The Food and Drug Administration (FDA) and the European Medicines Agency (EMA) are currently developing MCDA tools and processes for the benefit-risk assessment of new medicinal products.¹¹ The objective is to support informed, science-based, regulatory decision-making about new products. Other applications of MCDA are being explored in literature. Besides for the value assessment of health interventions for public coverage decisions and authorization decisions, MCDA could also be used for prescription decisions.¹¹

The current study builds upon the framework developed in KCE Report 147 and defines a set of weights for criteria for three of the questions (therapeutic need, societal need and preparedness to pay for the specific treatment) based on measurements in the general public. The question about preparedness to pay “in principle” (e.g. are we, as a society, in principle prepared to pay for treating a self-inflicted disease) is a fundamental ethical value-discussion, which was beyond the scope of the study. The question about preparedness to pay more is largely question of budgetary possibilities. These are by definition variable. As for cost-effectiveness ratios, it is complicated, if not impossible, to explain in a written survey the implications of choosing an intervention with a particular budgetary impact. It is therefore also not included in the current study.

The application of the MCDA principles to each of the questions considered will lead to three weighted scores. For reimbursement to be of high priority, all three weighted scores need to be high, or, if one of the scores is not high, there should be explicit reasons for still granting reimbursement. The actual reimbursement decision is thus a higher-level weighting matter: how much weight is given to therapeutic need, as compared to societal need and added value. Our study does not provide these higher-level weights for the different aspects of a new treatment and its targeted condition.

The current study provides one piece of necessary input for an MCDA tool that could be applied in a reimbursement decision-making context. The input provided is the weights to be applied to the scores for each of the attributes. The scores themselves are disease- and intervention-specific and need to be generated by the appraisal committees for each new decision. This ‘scoring’ is beyond the scope of the current study. The criteria weights indicate to what extent a criterion should be taken into account in the decision-making process.

The UK Department for Communities and Local Government (2009) developed a guide for MCDA which extends beyond health care decisions.¹³ We use this Table 7, derived from this guide and adapted to our study programme, to situate the current study within the broader context of the research programme that focusses on the application of MCDA in reimbursement decision processes and started with KCE report 147.² The first column of Table 7 gives the general guidance provided by the Department. The second column of Table 7 translates this guidance to the use of MCDA for health care reimbursement decision making. Trigger for starting the MCDA process is a reimbursement request for a new health technology.


Table 7 – General guidance for the application of MCDA and application to reimbursement decision making in health care

General MCDA Guide	Application in Belgian health care decision making
1. Establish the decision context <ul style="list-style-type: none"> Establish aims of the MCDA, and identify decision makers and other key players Design the socio-technical system for conducting the MCDA Consider the context of the appraisal 	<p><i>The aims of the MCDA, the framework for conducting the MCDA and the appraisal context has been studied in KCE Report 147.²</i></p> <p><i>The aim of the MCDA is to improve the transparency and consistency of reimbursement decision making and to incorporate public preferences with regard to reimbursement decision criteria. The framework is presented in Table 6 and originates from Report 147 that described extensively the Belgian appraisal context.</i></p>
2. Identify options to be appraised	<p><i>The identification of options to be appraised is a practical step to be taken during the reimbursement decision-making process. In a health care reimbursement context the options are the health interventions considered for reimbursement, but also target population and disease needs to be identified, as they are the basis for the appraisal of the therapeutic and societal need.</i></p>
1. Identify objectives and criteria <ul style="list-style-type: none"> Identify criteria for assessing the consequences of each option Organize the criteria by clustering them under high-level and lower-level objectives in a hierarchy 	<p><i>The objectives of the MCDA and criteria to be used in the MCDA are studied in the current study. In fact criteria are clustered in three hierarchical domains, each domain requiring a separate MCDA.</i></p>
2. Scoring. Assess the expected performance of each option against the criteria. To assess the value associated with the consequences of each option for each criterion <ul style="list-style-type: none"> Describe the consequences of each option Score the options on the criteria Check the consistency of the scores on each criterion 	<p><i>Scoring rules will be developed in a future study.</i></p>
3. Weighting. Assign weights for each of the criteria to reflect their relative importance to the decision	<p><i>The derivation of weights is part of the objectives of the current study. For each criterion within each domain (therapeutic need, societal need and added value) a level-independent criterion weight will be determined.</i></p>
4. Combine the weights and scores for each option to derive an overall weighted score for each level in the hierarchy	<p><i>Combining the weights derived from the current study with the scores as determined following rules that will be developed in a future study, is a</i></p>



	<i>practical step to be taken during the reimbursement decision-making process. For the future study, it is planned to pilot test this practical application, first for the aspect of therapeutic and societal need only, later also for added value.</i>
5. Examine the results	<i>This is part of the appraisal process.</i> <i>Weighted scores should be ranked. For therapeutic and societal need the weighted scores are compared with conditions assessed in the past using the MCDA tool. For added value, the weighted score is compared with interventions assessed in the past using the MCDA tool.</i>
6. Sensitivity analysis <ul style="list-style-type: none">○ Conduct a sensitivity analysis: do other preferences or weights affect the overall ordering of the options?○ Look at the advantages and disadvantages of selected options, and compare pairs of options.○ Create possible new options that might be better than those originally considered.○ Repeat the above steps until a requisite model is obtained.	<i>It should be tested in future research to what extent changes in preferences would affect the ranking of interventions on therapeutic need, societal need and added value.</i>

Note: All text in italics is our own interpretation of how the general guidance can be translated to reimbursement decisions in health care, not the authors' of the original guidance for MCDA.

Key points

- **The current study tries to identify possible rationales for decisions based on the preferences of the general public. Reasonableness is considered to be an operational goal of decision-making, as is transparency of the decision process. The study aims at supporting both goals.**
- **Multi-criteria decision analysis (MCDA) is a tool that could support decision-making, by helping to structure decisions and deliberative processes that involve multiple criteria. The current study tries to develop such a tool for making reimbursement decisions about new health interventions in the Belgian context.**



2 OBJECTIVES

The study fits within a broad research theme looking for ways to incorporate public values and preferences in decision-making processes in health care. The current study fits within a series of studies that all work towards the ultimate goal of developing a supportive decision tool that can help decision makers in making their reimbursement decisions more transparent and ensure that they are in line with the general public's preferences. The tool will not replace the decision process, but can, within its specific scope, help to increase the legitimacy of decisions. Before the tool can be developed, weights for the criteria that are identified as relevant need to be determined.

The objective of the current study is to provide weights for a set of generic reimbursement decision criteria, reflecting the relative importance of each criterion according to the general public. A general canvas for a future MCDA tool will also be developed. A future study will work on the next steps towards a workable multi-criteria decision analytic tool, using the weights determined in the current study as inputs. The **main research question** of the current study is therefore:

“What is the relative importance of reimbursement decision criteria according to the general public in Belgium and how can this information be used in a MCDA framework?”

A secondary objective of the study is to compare the weights for the same generic criteria derived from health policy makers with those derived from the general public. Such evidence will allow to explore whether the use of public preferences would change the actual decisions. It should be noted, however, that having an impact on the actual decisions is not the major or only impact this study aims at. The possible impact of using public preferences on the transparency and justification of decisions is of importance. An analysis of whether the use of an MCDA tool based on public preferences would actually change the decisions, their transparency and/or their justification, is not feasible within the time constraints of the current study and could be an objective of future research.

KCE and the King Baudouin Foundation (KBF) collaborate to deliver results that are practical, useful and validated. The ultimate objective of both organisations is to get a sufficiently differentiated idea of the public values and preferences with respect to reimbursement decisions in health care. KCE applies a quantitative approach for its analysis, while KBF applies a

qualitative approach, necessary to really understand what is behind the quantitative results or to put them in perspective. The KCE study will be followed by a “citizens’ lab” organised by the KBF, where citizens are informed about the purpose and design of the survey. The objectives of the citizens’ lab are (1) to explore whether deliberation by citizens leads to additional decision criteria, besides those put forward by committee members and experts, (2) to reveal arguments and considerations made by citizens when defining decision criteria and (3) to assess the ranking of importance of criteria after deliberation. The methods and results of the citizens’ laboratory will be reported in a separate report written by the KBF. The current report describes the results of the quantitative part developed by KCE.

Key points

- **The main objective of the current study is to elicit preferences for reimbursement decision criteria from the general public.**
- **A secondary objective is to compare the relative preferences for reimbursement decision criteria from the current policy makers with those of the general public. This evidence could be used in future research to explore whether the use of a MCDA tool changes the actual decisions, and whether it improves the transparency and justification of the decisions.**
- **A qualitative research project is set up by the King Baudouin Foundation in collaboration with KCE to further dig into the results of the current quantitative study in order to better understand what is behind these preferences.**



3 GENERAL METHODOLOGICAL APPROACH

For the development of an MCDA framework, first, the criteria for making decisions need to be defined. Possibly relevant criteria were identified based on a literature review. Next, the relative weight of each criterion considered relevant, needs to be determined. Our objective was to measure the preference weights assigned to the criteria by the general public. Therefore, we decided to perform a large population survey. For this survey, we had to choose a preference elicitation technique. Several techniques have been described in literature. We performed a literature review of preference elicitation techniques and organised a workshop with external experts to be able to make an informed choice.

The eventually selected preference elicitation technique (discrete choice experiments) required a reduction in the number of criteria identified as possible relevant, as too many criteria would make the choice task unfeasible for respondents to our survey. The selection of the criteria to be included in the survey was done based on discussions with groups of experts.

The methods and results of the literature review and the selection of a preference elicitation technique and criteria are presented in Chapter 4.

Chapter 5 presents the methods and results of the population survey.

4 LITERATURE REVIEW

4.1 Methods

A literature review was performed to identify possibly relevant decision criteria to be included in our survey and possible measurement techniques for preferences regarding decision criteria. This review was then discussed by an expert group, consisting of internal KCE and external experts with different types of expertise (see 4.1.5).

4.1.1 Literature search strategy

The databases Medline, Embase and Sociological Abstracts were searched. The search strings are presented in Table 8. In addition, a hand-search was performed to include other relevant papers in the grey literature, for instance for identifying examples of MCDA frameworks which are either already applied in real life or are the output of a collaboration between researchers from different countries (such as the EVIDEM collaboration). The search was performed on 19 August 2013 (Medline and Embase) and 22 August 2013 (sociological abstracts) and has not been updated since because of time constraints.

Table 8 – Search strings

Database	Date search	of	Search string	Hits
Medline (PubMed)	19/08/2013		("Social Values"[Mesh]) AND "Health Priorities"[Mesh]	176
Embase (OVID)	19/08/2013		"social values" and "health priorities" AND ([article]/lim OR [review]/lim) AND ([dutch]/lim OR [english]/lim OR [french]/lim OR [german]/lim) AND [embase]/lim:	305
Sociological Abstracts	22/08/2013		"su.exact("social values") and "health priorities"	5 (1 relevant)



The snowballing technique was used to identify additional relevant references that were not picked up by any of the search strings but are essential for our review.

4.1.2 Scope of the literature review

The scope of this literature review is limited to literature about priority-setting criteria in health care in a national or regional context applicable to different types of services. Issues of patient selection for treatment in a health care practice setting are beyond the scope of this study.

The focus of the search strategy was mainly on empirical studies applying a specific preference elicitation technique or multiple techniques to learn about these techniques and make a decision on which technique to use for our survey. We tried to learn maximally from existing multi-criteria decision frameworks and empirical research concerning societal preferences for priority-setting criteria in health care.

4.1.3 Selection of studies

Inclusion and exclusion criteria are presented in Table 9. The inclusion and exclusion criteria related to the scope of the studies, population and methods directly relate to the objectives of our literature review. As our objective is not to create a priority list of concrete health interventions, but rather to develop a framework that can help to develop such a list based on generically defined decision criteria, we excluded papers that did not describe health interventions by means of explicit pre-defined criteria. For example, diabetes care can be compared with dialysis, neonatal care, breast cancer screening etc. using narrative descriptions of these diseases without a standard structure or with a standard structure that is based on different relevant criteria. The structured approach is the one relevant for our work. Although empirical studies in populations other than the general public or citizens could provide useful insights, we limited our search to studies performed in citizens. The literature review was not a deliverable as such but had to provide inputs for our own empirical study. We therefore focussed on aspects that were of direct relevance to our study.

Studies applying only qualitative techniques for eliciting public preferences or narrative descriptions were excluded from the review. Techniques included in the review had to allow for the estimation of the relative strength of preference for several decision criteria or attributes of diseases, patient groups or interventions.

Our primary objective is to get insight into the relative importance of operational criteria relating to health interventions, individuals and conditions according to the general public, but we realize it is impossible to consider these completely separately from the guiding ethical principles for health care resource allocation decisions. Guiding ethical principles for health care resource allocation refer to social norms, i.e. what ought to be done, how operational criteria should be handled to lead to a fair and just allocation of health care resources. It is recognized that (the sum of) *individual preferences* for handling operational criteria might not coincide with *societal norms*. Most often it is individual preferences rather than norms that are studied in empirical literature, because societal norms are difficult to measure, but studies that defined the measurement of societal norms as their primary objective were not excluded from the review.

**Table 9 – Inclusion and exclusion criteria for the literature review**

	Inclusion criteria	Exclusion criteria
Scope of health care priority setting	Priority setting for defining the general health care benefit package or reimbursement of health care interventions, focus on multiple “content” criteria. Also structured narrative health intervention descriptions are allowed, if criteria are clearly identifiable.	Focus on process criteria, prioritization of patients, prioritization within a particular health care facility, selection of treatment for a particular disease, definition of a specific health care program or intervention, measurement of health-related quality of life (HRQoL) in diverse health states (e.g. for defining the HRQoL weights in Quality-Adjusted Life Years (QALYs))
Population	Citizens, general population	Subgroups of the general population (students, policy-makers, healthcare professionals)
Methods	Quantitative methods (e.g. ranking, rating, contingent valuation) or a combination of qualitative and quantitative methods	Qualitative methods only; narrative descriptions of health interventions without standardized structure
Language	English, Dutch, French, German	Other languages
Publication type	Journal articles	Letters, editorials, opinion papers, commentaries

4.1.4 Data extraction methodology

A categorization scheme was developed to classify the criteria used in literature. We made a distinction between two categories of criteria: (1) criteria relating to general overarching ethical principles and (2) criteria relating to the characteristics of interventions, individuals or health conditions. The scheme was used to structure the evidence tables for the empirical studies identified.

The advantages and disadvantages of different preference elicitation techniques were described in evidence tables. In order to have an overview of the strengths and weaknesses of different methods, the studies with a similar methodology were pooled into groups. Studies employing more than one method, were categorized according to their main focus. For example, if a ranking and a rating method were used, but the rating was merely a check for the ranking, than the study was classified as a ranking study.

The methods were evaluated using the same evaluation criteria as in the systematic review of preference elicitation techniques of Ryan et al. (2001)¹²: validity, reproducibility, internal consistency and acceptability.

The **validity** criterion looks into the goal and the outcome of the method and questions whether the adopted method measures what the researcher is trying to measure. Validity focuses on the relation between the questions and the results. Theoretical validity can be tested by assessing to which extent the results are consistent with a priori expectations. The validity of the method can also be checked by questioning the same respondents with two different techniques and compare the results. Framing or wording of questions and experimental conditions might affect the validity of a method. It is very hard, though, to assess the validity of the framing, as it is unclear what the golden standard is. Evidence shows that the framing of questions plays a very important role in finding preferences. Nord et al.(1996) compared a direct approach, asking whether younger people with life threatening illnesses should get some priority over older people with life threatening illnesses or whether they should get the same priority, with an indirect approach (person trade-off). The results were quite opposite. In the direct approach, 41.9% would give the same priority to the young and the old with respect to life saving treatment, whereas in the indirect approach there was a strong preference for giving priority to the young, even if both would have the same life expectancy.¹⁴ The authors describe several



possible reasons for this discrepancy, including the effect of question framing. The direct question was framed more personally, whereas the person trade-off question was framed impersonally as “in a budgeting context”, thereby influencing the responses of the participants.

In research that tries to investigate preferences for priority setting, it is tricky to just ask respondents for their opinion, because people might not really know what their opinion is about the topic. It is possible that people are confronted with questions they would never think of themselves. Thus the type of question becomes crucial. Ranking exercises make people implicitly conscious about their arguments and the relative weight of their arguments.

Acceptability refers to the extent people accept the type of questions asked and the method used to collect the data. Acceptability focuses on the feeling of the respondent about the questions. For example, some people might find questions about which of two patient groups they would prefer to be treated especially unacceptable or threatening. Acceptability can be measured by asking the respondents whether the questions are not too hard or too unrealistic, or by measuring the completion time or completion rate. In this context, the discussion arises about the use and interpretation of “non-response”.¹⁵ This could leave a choice for people who do not understand the question or who do not want to answer because they think the public should not decide upon this issue or other reasons. Data collection techniques and experimental conditions also influence the acceptability.

It is important to distinguish between difficulties resulting from the content of questions and difficulties arising from the framing of the questions. Difficulties due to the content are not necessarily a disadvantage of a specific method, because some topics are more sensitive than others. Setting priorities in health care creates hard ethical choices. However, there is still a margin to make these choices more or less complicated, by giving more or less information to the respondent, for instance. Often scenarios allowing many variables to change at once in every choice case are more difficult to answer. Framing questions easily, on the other hand, does not necessarily give valid responses, because they could omit important considerations determining respondents’ actual values.

Some studies use more advanced programmes or statistical models to construct and analyse the results to correct for the failure of people to explicitly state their preference. This helps the deduction of preferences in, for instance, conjoint analysis, where multiple criteria are compared with

each other. However, when using statistical models to estimate preferences, assumptions have to be made. Rarely information about the underlying assumptions is presented, creating a loss of transparency of the applied approach.

Internal consistency assesses the logic of respondents’ answers. This criterion assumes that people use a certain consistent logic in line with the cognitivist meta-ethical assumptions of rationality. This is often tested with a dominant option or with a transitivity test. The dominance test assesses whether an option that clearly dominates another because it is better in all respects is actually preferred. If not, the respondent’s answer is inconsistent. For the transitivity test, three options are presented in different combinations. If one option is preferred above the other option and that option is preferred above another option, then the first option should be preferred above the last option (if $A > B$ and $B > C$ then $A > C$). Transitivity is one of the axioms applied in choice theory in economics to describe preference orderings. The other axioms are reflexivity and completeness of preferences. If an individual’s preferences are reflexive, transitive and complete, a preference ordering exists.¹⁶ Reflexivity is less interesting from a practical point of view as it only refers to the fact that a specific scenario is at least as preferred as itself. **Completeness** means that individuals always have a preference about two alternatives. The preference can be that one is preferred over the other or that both are equally preferred. Completeness is rarely tested because it is very difficult to test in a valid manner. Attempts, such as the study by Shiell et al. (2000)¹⁶, have been criticised.¹⁷ It is argued that people participating in surveys generally attempt to appear consistent. They might therefore be sensitive to such consistency test questions. Ryan et al. (2001)¹² did not include completeness as an evaluation criterion in their review of preference elicitation techniques.

Reliability can be defined as reproducibility of the results over time. Usually this is tested by repeating the same exercise over a short period of time and then comparing the results. However it should be noted that this criterion assumes that preferences are stable over time, which might not necessarily be the case. Moreover, it can be questioned whether we really need reliability on the individual level for our purposes. More important is stability of the results on the aggregate level, as this is the level at which the data will be used.



Not covered in these criteria but nevertheless relevant for any empirical study is the method of aggregating individual responses. Wiseman et al. (2004) found that the aggregation method influences the final rankings of programmes or criteria.¹⁸ The number of options presented to respondents also influenced the decision to allocate funds equally. When there are only two options, the vast majority of respondents give equal allocations, whereas only 16% gave equal allocations when four options were presented.¹⁸

4.1.5 External expert group discussions

The literature review served as a basis for a discussion with external expert groups. Different groups were established for different aspects of the study: one for discussing the criteria to be included in the survey, and one for discussing the preference elicitation techniques. This allowed us to target the discussions and to limit the group, which stimulates active participation of all experts.

The groups convened on a different date and had a different composition. The **criteria to be included** in the survey were discussed with an external expert group during a half day workshop on 26 September 2013. The group consisted of experts with different backgrounds, mainly academics: sociology (with extensive expertise in survey research), public health, philosophy, communication towards lay public (journalism), biostatistics and biomedical research (with extensive expertise in multi-criteria decision analysis).

The experts received preparatory documents in advance of the workshop: a document describing the background and objectives of the study, a brief description of the methods, a long-list of possible decision criteria based on our literature research and a description of a number of existing transparent decision-making frameworks (see appendix). The description of a number of MCDA frameworks currently used in practice was based on a recently published extensive systematic literature review. No separate systematic literature search was performed for this purpose. The objective of presenting existing MCDA frameworks was to give the experts an idea of how such a framework could look like in practice, rather than to give a full systematic overview of all existing frameworks. The description should help the experts to understand the concrete purpose of their task. Each framework contained a set of explicit decision criteria, which was completed with the criteria

identified in the literature review. The research team prepared a summary of criteria that fitted within three out of five questions of the KCE MCDA framework presented in Table 6.

The objective of the workshop was to define a list of possibly relevant reimbursement decision criteria for Belgium that satisfied the requirements for multi-criteria decision analysis, being:⁸

- clearly defined and based on clearly articulated principles;
- operationalisable, i.e. it must be possible to describe or measure the characteristics of the options that decision makers are considering in terms of these criteria; and
- orthogonal, i.e. they should not just be alternative measures or proxies of the same underlying principle: each criterion covers one and just one dimension of potential interest.

The summary prepared by the research team was used as a starting point and was complemented with additional criteria if deemed necessary by the expert group. The summary of possible criteria fitting into the 5-question framework was adapted on the spot during the discussion by the principal investigator. The document in which the adaptations were made was projected on individual screens, so that every expert of the group could see which modifications were made following the discussion.

Following the workshop, the research team created a clean draft list of criteria which was re-discussed within the team to check the consistency of the new list with the criteria for MCDA and to discuss possible ways to operationalize these criteria. It is important that criteria can be operationalised in a way that is comprehensible for all citizens.

Moreover, it was clear that already in the selection of the criteria, choices have to be made. For some criteria, the decision was particularly difficult. "Age", for example, was not considered as a relevant criterion *per se* during the expert meeting. However, it was considered *possibly relevant* in relation with impact of a disease on life expectancy. For example, dying immediately from an illness might be judged differently for an 85-year old than for an 18-year old by the population. A decision maker might not want to take this into account, but excluding age from the generic description of a case might lead to too many blank responses from people who consider age important for the decision, not because of the criterion *per se* but rather for the potential



benefits to be gained. When a young patient, who would otherwise die immediately from his disease, could be saved, the duration of his expected benefits is longer than when an older patient would be saved. The final choices made by the research team were discussed with the external experts group through e-mail. Several e-mail exchanges with the external experts led to the final list of attributes with their different levels to be used in the survey.

The choice of the **preference elicitation technique** for the survey was based on discussions with an external expert group consisting of experts in the field of preference elicitation with empirical experience. Two health economists, a biostatistician, a sociologist, and a health service researcher shared their personal experience with a specific technique or several techniques. Different options for the Belgian survey were discussed, using the summary of an existing literature review of techniques as a basis. No final conclusion about the most appropriate technique was reached during the workshop, because all techniques have strengths and weaknesses.

The discussion about the preference elicitation technique to use in the survey continued after the selection of the relevant decision criteria to be included in the survey. The internal research team re-examined the different techniques in the light of the objective of obtaining relative preferences to be applied in MCDA and proposed a short list to the accompanying external experts for this project, Mrs Janine van Til and Mrs Karin Groothuis. Both have a broad experience with several preference elicitation techniques and shared their thoughts about the optimal technique. This exchange of thoughts reduced the list of techniques to three appropriate techniques for our purposes: discrete choice experiments, best-worst scaling or “ranking and rating”. The eventual choice of a technique was made by the KCE research team.

Key points

- **A literature review was performed to identify and describe empirical studies about societal preferences for setting priorities in health care. This helped to identify possibly relevant decision criteria.**
- **An external expert panel, gathered in one face-to-face workshop and consulted by email afterwards, allowed to reduce the long list of criteria to a manageable short list.**

- **The literature review also identified possible techniques for eliciting public preferences for reimbursement criteria.**

4.2 Results: Criteria for making reimbursement decisions

4.2.1 Flow-chart literature review

The flow chart of the literature search strategy is presented in Figure 5.

Figure 5 – Flow chart literature search





4.2.2 *Classification of principles for rational resource use and choice criteria*

Reviewing the literature on societal preferences for resource allocation in healthcare is challenging, as study findings regarding specific criteria are conditional upon the other criteria included in the study. We attempted to summarize the elements for which the empirical evidence was rather consistent as well as those for which the evidence was not consistent. The review does not try to explain why studies reached certain findings.

In the review, a distinction is made between evidence regarding general ethical principles for making rational resource allocation decisions in health care, and operational intervention-, patient- or condition-related criteria.

General resource allocation principles include (1) the lottery principle or “not playing God”, (2) the rule of rescue or distribution of resources according to immediate need, (3) health maximisation, (4) fair innings or equalizing lifetime health, and (5) choicism or equalizing opportunity for health.^{19, 20} They either imply that one (combination of) operational criteria is dominant to all other (combinations of) operational criteria (e.g. in case of the rule of rescue, a life-saving treatment for an immediately life-threatening unmet medical need dominates everything else) or that person-, intervention- or condition-related criteria are irrelevant (e.g. in case of the lottery principle). Operational criteria should be compatible with more general ethical principles, but must be developed in their own right.⁶

Patient-related criteria encompass age, social worth, socioeconomic status and ability to benefit.

Condition-related criteria encompass therapeutic need, severity of disease (medical need) and disease frequency.

Intervention-related criteria encompass the intervention's safety, efficacy, effectiveness in terms of life extension and quality of life improvement, cost, level of uncertainty with respect to effectiveness of treatment, its level of innovativeness (also inducing potential future innovations), and the type of treatment (prevention or treatment, care or cure).

Besides the general ethical principles and operational criteria for reimbursement decision making, there are also procedural guidelines. These are beyond the scope of the current project but were developed in a previous KCE report.²

A summary table, including the criteria included and the results and conclusions of the empirical studies, is presented in appendix. The overview of the criteria was complicated by the fact that different names are often used for similar abstract concepts. Moreover, many studies do not explain what they mean by the criteria. Therefore, our review only summarizes the results as presented in the studies, involving the value judgments regarding the underlying interpretation of the criteria as made in the studies.

4.2.3 *General principles for resource allocation*

Possible principles for rational allocation of resources in health care include:

- the lottery principle;
- the rule of rescue;
- health maximisation;
- fair innings;
- choicism.

In a survey amongst Thai citizens investigating the preference for these five rationing principles, it was found that all principles were used, depending on the specific decision problems presented, but choicism came out most frequently as the preferred principle.²⁰ Decision problems were formulated as choices to be made between two patients with the same severe disease but with different levels of pain, different health gains from treatment, different waiting times until treatment, different ages and life expectancy or different causes of disease (drug abuse or bad luck). Choicism gives priority to those who suffer from diseases that are not a result of patients' own lifestyles.

Fair innings came out as preferred principle in three out of four choice problems that included patients of varying ages. The rule of rescue was preferred to health maximisation and lottery and health maximisation was only preferred to lottery. The lottery principle (“first come first served”) was never preferred.²⁰ Preferences for resource allocation principles obviously depended on the specific situation presented to the respondents, and not all situations could be presented in that single survey. However, the results do provide some insight into the most preferred principles.



The evidence suggests that it is often not possible to identify one single principle that should always be applied. Most people prefer a combination of needs-based, health maximising and egalitarian principles.²¹

4.2.3.1 Lottery principle

The lottery principle is based on the principle that every life is of equal value and hence one cannot choose between patients or interventions based on explicit criteria.²² If the lottery principle is accepted, everyone should get equal chances for treatment. In choice experiments, this would lead to respondents not being willing to choose between scenarios or stating that all patients should get an equal amount of resources. In clinical practice, treatment would be given on a first come first served basis or to those who are on the waiting list for the longest time. However, in a coverage context, it is difficult to understand how the lottery principle could work. Would it imply that equal allocation of resources to all is considered more important than characteristics of patients, health conditions, (potential) health benefits or interventions?

Despite the evidence that this principle is never preferred if compared with other resource allocation principles,²⁰ evidence exists that the general public does support this principle but not as the only and absolute principle.²³⁻²⁷

In studies where people were given the opportunity to opt out of the choice decision, up to 50% did not want to be involved in such a decision process.²⁵ This result contrasts with the results from Green et al. (2009), where only 5% took the option to opt out.²³ People who opt out could be considered to rather prefer the lottery principle than to make a choice based on patient criteria. However, evidence also exists that if respondents are given the opportunity to opt out of the choice problem, they tend to do so, but if they are not given the chance to opt out, they would prefer to allocate more resources to the severely ill even if they would benefit less from treatment than others.²⁴ Anderson found that many respondents to his survey were strong egalitarians, giving equal priority to all patients, although few (5%) were consistent egalitarians. Most people were prepared to discriminate in a limited number of cases.²⁸

As an alternative to the lottery principle, people might choose to allocate an equal amount of resources to all patients.^{24, 25} This is another type of opting out of the choice problem and not having to apply the lottery principle. Related to this, Wiseman et al. (2004) found that the number of scenarios

between which respondents have to choose in a choice experiment determines the extent to which they want equal allocation of resources. The vast majority of respondents in their study gave equal allocations when only two options were presented, whereas only 16% gave equal allocations when four options were presented.¹⁸ Hence, the framing of the questions (type of questions, response options given and number of options to choose from) all determine the extent to which support for the lottery principle or for equal allocation of resources is found.

The equal allocation of resources is not equivalent to the lottery principle but often considered to be a possible way out of having to make a choice. This is, however, a pitfall. Giving an equal amount of resources to two patients either means that the patient has to cover the cost of the remaining resources needed to treat his/her disease or both patients can only be treated incompletely. Both results of an equal allocation of resources are not without moral consequences, especially if the allocation is made without information on the patient's financial situation (most relevant for the first possible result of the choice).

4.2.3.2 Rule of rescue

The rule of rescue principle refers to the allocation of resources according to the immediate need. Empirical evidence shows public support for this resource allocation principle.²⁹⁻³⁵ Most studies find that people are willing to give higher priority to treatments for life-threatening conditions than to treatments for less serious illnesses.³¹

A strong support for the rule of rescue was found by Zweibel et al. (1993), showing that few respondents would categorically withhold life-prolonging medical care to critically ill persons who are near death, even if they are old and unlikely to recover.²⁹ Interestingly, they also found that the older respondents consider extending the lives of dying elderly as wasteful.²⁹

Also Oregon's first priority list for health care in 1990 was heavily criticised precisely because life-saving treatments were not systematically ranked higher than minor, but more cost-effective treatments. It was felt that the rule of rescue was in this way overruled by the principle of health maximisation.³⁶ The final list of priorities in health care was established without reference to cost-effectiveness.



Shmueli et al. (1999) found strong support for the rule of rescue in Israel.³⁷ In a comparison between the relative importance of rescuing life (or prolonging survival) and preventing severe and permanent disability, they found that 27% of the respondents attaches a high value to the act of rescuing human life, even when death is postponed by only one month. Rescuing life is preferred over preventing a dramatic decline in the quality of life. One rescued year of life with a certain degree of dependence for a patient with a life-threatening condition was valued higher than 30 years of remaining life in severe disability of a patient with a non-life-threatening condition (40% of respondents). The marginal value of a life year saved diminished, however, with an increasing survival period, meaning that gaining more life years after the rescue is still more valuable than gaining less life years, but the increase in the value is not proportional to the increase in life years.

The relative value of rescuing life as compared to other benefits deserves further attention. In a longitudinal study, involving four repeated surveys in the general public in the period between 1997 and 2004, Chinitz et al. (2009) found that over time there seems to have been a shift from prioritization of life-extending treatments towards increased relative preference for treatments adding quality of life.³⁸ Other evidence also finds that people think patients' quality of life should also be taken into account when deciding to cover a lifesaving intervention.^{30, 34} Another study found that the number of lives saved is the most important decision criterion (weight 0.343), although it should be weighed against other criteria, such as life-prolongation benefits (weight 0.243), quality of life gains (weight 0.217), availability of alternative treatments (0.107) and other social/ethical benefits (0.087).²¹ Similar conclusions can be drawn from other studies: not only the type of the health gain is important, but also the size, who receives the health gain and sometimes whether the individual is responsible for his/her own health state.³³

4.2.3.3 Health maximisation

Health maximisation originates from the philosophy applied in classical health economics that resources should be allocated in such a way that the total health of the population is maximised. Resource allocation based on incremental cost-effectiveness ratios uses health maximisation as a guiding principle. The principal objective is then to maximise the number of Quality-

Adjusted Life Years (QALYs) gained, regardless of who gains these QALYs or how many people gain QALYs.^{23, 39}

Most studies find that societal values are inconsistent with simple health maximisation.^{23, 39, 40} The general public is generally willing to deviate from the health maximisation principle in favour of giving priority to the more severely ill.^{23, 39} People prefer less costly and more effective interventions when setting priorities, but take non-health arguments (such as age, curative or preventive intervention, strength of the evidence, and personal contribution to the illness) into account.^{25, 39} One study found that, when asked directly about the importance of allocation principles, 35% of the respondents considered that treatment outcome of the recipients was the most important factor for choosing between potential beneficiaries of high-cost medications, followed by current health status (26%) and quality of life (15%).²⁵ Another study also found that the majority of the public (70-80%) might prefer geographic equality in the distribution of health gain over health maximisation.⁴¹ However, for the majority of those who trade-off health for geographic equality, the sacrifice in terms of health should not become larger than 10%. Interestingly, respondents were also prepared to give up vaccination for 10 children out of 100 in order to be able to vaccinate an equal number of children in two regions. The number of patients undergoing surgery given up for the sake of geographic equality of health distribution was even higher.⁴¹

4.2.3.4 Fair innings: equalizing lifetime health

The fair innings, or equalizing lifetime health, principle is based on the idea that everyone in the society should be allowed to reach a certain amount of lifetime health. It aims at minimising health inequalities among people. The fair innings principle favours the younger and disabled people, because their expected lifetime QALYs is lower than the old/non-disabled.⁴² Hence, fair innings is related to age, though not exclusively. Also people's past health state determines the extent to which they have lived their fair innings.

Dolan & Tsuchiya (2005) found that there is a strong effect of age: younger groups (40-year olds) are always chosen over older ones (60-year olds). In addition, patients with worse past health are more likely to be given priority than those with good past health,^{27, 43} even if their health gains are smaller.²⁷ Future health and future years without treatment, on the other hand, are



sometimes found to be non-significant,⁴³ while other studies find remaining life to be important considerations in priority setting.²⁷

Similarly, Lees et al. (2002) found that people with chronic illnesses, people with physical disabilities, children, people who are mentally ill, people living in poverty and people who are terminally ill should receive a higher priority for health care, but also the elderly.³¹ The latter finding seems to be in contradiction with the fair innings argument, while the former findings seem to support the argument.

4.2.3.5 *Choicism: equalizing the opportunity for health*

Choicism refers to prioritization based on “bad luck” versus own responsibility for illness. The evidence suggests that people are willing to discriminate based on lifestyle-induced diseases. For example, when asked whether a drunk driver should bear the costs of medical care that he needs after a careless car accident he caused, respondents generally answer “yes”.⁴⁴ However, empirical studies revealed differentiated responses. For example, Fowler et al. (1994) found that patients ranked coverage of treatment for a no-fault accident victim higher than coverage of the treatment for a driver who drove too fast. However, the proportion of respondents who nevertheless found coverage was needed for both accident victims was still large (97% versus 88%).⁴⁵

There is a large literature on the relative importance the general public wishes to attach to personal responsibility for making resource allocation decisions in health care. Personal responsibility is related to a perception of accountability. It assumes that individuals have a certain power to influence the adverse events in their life regarding illness.

A recent survey in the Belgian general public found support for letting patients contribute financially to the health care system in function of their lifestyle: 21% of the respondents found that people should be tested every year for physical condition and those who are fit should contribute less to public health insurance. For smoking and alcohol use the percentage of respondents in favour of lower contributions for those who do not smoke or drink alcohol are respectively 25% and 28%.⁴⁶ Besides differentiation in contributions for social health insurance, the study asked questions about reimbursement of treatments for self-inflicted conditions. Between 36.5% and 50% of respondents found that people should get reimbursement, despite reckless behaviour or unhealthy lifestyle, depending on the case

presented. 18% to 27% would rather not reimburse treatments in the same way as for people with healthy lifestyles or cautious behaviour. The remainder of the respondents was undetermined.⁴⁶ The results show that the preference for discriminating based on lifestyle factors is not absolute: although a considerable percentage would support differentiated reimbursement (especially if it concerns “others”, not “themselves or their family”), still a large percentage of people would prefer not to differentiate.

Another Belgian survey, using discrete choice experiments, also found that own responsibility for illness is an important factor for the public when choosing which interventions to give priority for reimbursement.⁴⁷ Amongst the attributes lifestyle of patient, age of patient, effectiveness, severity of illness, adverse effects, timespan of effects and type of intervention (prevention or cure), the lifestyle of the patient was the most important attribute, meaning that people would rather not give priority to interventions for self-inflicted diseases.⁴⁷

The World Health Organization (<http://www.who.int/hia/evidence/doh/en/index.html>) states that the determinants of health include income, social status, social support networks, education, physical environment, genetics, gender and access and use of health services. The belief that all of these elements, besides gender and genetics, can be influenced by the individual and thus that the individual has a responsibility towards his or her own health status,⁴⁸ could be an explanation for some of the results observed in literature.

Lifestyle is often used as proxy for personal responsibility. For example, smokers, people with unhealthy diets, people taking drugs, who rarely exercise and who over-consume alcohol may be perceived as self-harmers.²⁸

The results with respect to personal responsibility as a criterion for resource allocation are mixed. Some studies find some support for choicism based on individual responsibility,^{30, 34, 39, 49} whilst other studies find that lifestyle has a negligible weight as prioritization criterion.^{28, 50}

In several studies, about half of the respondents support choicism, while the other half does not.^{30, 31, 34} For example, in one study 42% of the respondents agreed with the statement that people who contribute to their own illness, for example through smoking, eating, or excessive drinking, should have lower priority for their health care. 43% disagreed.³⁰ Young people seem to be less



in favour of using responsibility for own disease as a prioritization criterion.³⁴ People also seem to make a trade-off between personal responsibility and severity of disease: while people may a priori wish to give no support to a patient whose lifestyle was mainly responsible for a disease, they often change their initial decision when the case becomes more severe.²⁷

At the same time, there seems to be a trade-off between personal responsibility and cost-effectiveness: in cases of self-inflicted diseases, the acceptable incremental cost-effectiveness ratio seems to be lower.³⁹

Interestingly, Edlin et al. (2012)⁴⁸ believe there is a strong link between socioeconomic disadvantages and absence of responsibility for own illness. They surveyed people from the general public in the United Kingdom (UK) to assess the relative importance of responsibility for own health and inequality in lifetime health. They found that people wish to give less weight to blameworthiness when the patients also experience poorer health prospects. Having poor health prospects is weighted more heavily than being responsible for one's own health condition.

4.2.4 Patient-related criteria

4.2.4.1 Age

It is often assumed that the public prefers the young to be prioritized over the old.^{51, 52} Evidence supporting this assumption exists, but also highlights the conditional nature of these preferences.

Eight studies^{14, 26, 30, 32, 34, 39, 43, 48, 53} in our review showed that young people were preferred over old, and eight studies^{18, 27, 29, 31, 49-51, 54} showed no or conditional preferences for prioritization based on age.

Results regarding the acceptability of age-based coverage varied depending on the approach used to measure this attitude²⁹ and the other criteria included in the questions.^{30, 34} As for the criteria included in the scenarios, Bowling et al. (1996) and Mak et al. (2011) found that people agreed that high-cost technology should be available to all, while "in general" priority should be given to the younger.^{30, 34} The preference for giving priority to the young is also conditional upon the combination of length of life and quality of life. Considering length of life separately, people tend to give preference to the young, because their life years gained are likely to be higher. However, in combination with quality of life, the preference may change, especially if the young would live very long at a bad quality of life and the

old would live shorter at a good quality of life.²⁰ Many studies have not clearly taken this conditional preference into account when asking the public about age, life expectancy and quality of life.

Tsychiya et al. (2003) list three frequently cited reasons for prioritizing the young: health maximisation (longer life expectancy), productivity (more productive years ahead), and equity (the young have not yet had their fair share of life).⁵⁵ The observed preference for young children in stated preferences studies regarding saving lives may be attributable to the interactions between health and age that are generally ignored in such studies. Young age is usually associated in people's minds with longer life expectancy and better quality of life.^{20, 39} As described by Mak et al. (2011) "it is not easy to distinguish between age discrimination per se and prioritization based on other criteria that are associated with age".³⁴ Respondents expect interventions targeting young children to save more life years per life saved than interventions targeting the elderly. Age becomes a proxy for capacity to benefit.^{43, 56} For example, denying coronary artery bypass surgery to a frail 85-year old patient is justified because of the high risk of death during surgery. Denying treatment based on fewer expected quality of life-adjusted years may also be considered a valid reason for denying treatment to some elderly patients.³⁴ Nord et al. (1996) corrected for this possible bias by letting people compare health programmes with the same immediate impact on life expectancy (10 years), but targeted at different age groups.¹⁴ They still found a preference for treating the young, with preferences becoming stronger the larger the difference in age between the groups (i.e. giving ten additional life years to ten 20-year olds is considered equivalent to giving 10 additional life years to four 60-year olds or to one 80-year old).¹⁴

Age may also be a proxy for social worth. Diederich et al. (2012)⁵⁰ found that the most preferred age was 43, which represents people of working age, and utilities decrease for both decreasing and increasing age, with a steeper decrease for increasing age.⁵⁰ Mortimer et al. (2008) have demonstrated that in a model comparing life-years saved (instead of lives saved), interventions targeting young children and young adults are still preferred, but the preference is weaker.³⁹



A reduction in the strength of the preference for the young was also found in a study where respondents were asked to perform a moral exercise before answering the prioritization questions.²⁶ The moral exercise consisted of selecting 3 out of 10 possible allocation principles deemed most important for the scenario under study. The preferred principles did not reflect a preference for age-dependent health care coverage and the results of the resource allocation experiment also showed a lower importance of age as decision criterion.²⁶

4.2.4.2 *Social worth*

The relevance of having a family or being married is rather low in the resource allocation debate.^{28, 49, 50, 53} Gallego et al. (2007)²⁵ and Diederich et al. (2012)⁵⁰ found that family commitments are ranked at the bottom of the list of factors to be considered when deciding about the coverage of health care services, while in Johri et al. (2009)²⁶ they were ranked in the middle. If at all, married patients, patients in demanding caring roles in society for either children or elderly, sole breadwinners and good community contributors are given a higher priority for health care services.^{28, 53}

One study found that the public is strongly against giving priority to patients who hold important positions in society or are responsible for a family.⁴⁹

4.2.4.3 *Socioeconomic status*

The conclusions from the empirical evidence with respect to socioeconomic status are mixed. Several studies found that there is a preference to give priority to the more disadvantaged, all else equal^{18, 23, 27} whilst others found that compared to other criteria, socioeconomic status has only a minor or no role to play.^{50, 51, 53, 57} Most people find it impossible to choose between socioeconomic classes. For example, when asked to choose between occupational classes (managing director versus unskilled worker), most people responded they could not or did not want to choose. The people who did choose, gave slight priority to the unskilled worker.⁵³

One could expect that when medicines become more expensive that people would choose to take socioeconomic status into account, but Gallego et al. (2007)²⁵ found that socioeconomic status was ranked very poorly.

As for most other criteria, people make trade-offs between preferences for health gains, current health status or need and preferences for helping the worst off. For example, Tsuchiya et al. (2007)⁵⁷ found that around half of the

people preferred to spend resources evenly across social classes, but when the sacrifice became too big in respect of health gains of the better-off half of the people prioritized the better-off. These findings were not confirmed by another study, that found that even if they gain less health with the same treatment than others, lower socioeconomic classes were given priority over higher socioeconomic classes.²³ To further illustrate the impact of other criteria on the preference for giving priority to the socioeconomically disadvantaged, another study found that 26% of the respondents to a survey preferred to give priority to the disadvantaged socioeconomic groups when their health is worse, but 40% preferred to give an equal share to the advantaged and the disadvantaged group.⁵⁸ When besides current health of the two groups also criteria of efficiency and need were included, most respondents changed their preference of giving priority to the disadvantaged. Typically, higher priority was given to treating those with the highest needs first (whether advantaged or disadvantaged); only after this has been realised, participants turned back to their equality preferences as secondary considerations shaping the resource allocation.⁵⁸

4.2.4.4 *Ability to benefit and chances of success*

The outcome of an intervention is always uncertain to some extent. The “ability to benefit”-criterion is meant to elicit the weight people give to a conditional result of interventions. The ability to benefit expresses the chance of success of an intervention on a specific patient or patient group. In a health maximisation context, one would expect that people have a stronger preference for treatments with higher chances of success and would hence allocate resources primarily to patients with the highest capacity to benefit from the intervention. The evidence with respect to this criterion is limited.

The study by Gallego et al. (2007)²⁵ found that, in a hypothetical scenario where respondents have to choose between patients, 80% of the respondents favoured a choice based on ability to benefit in terms of quality of life and length of life.²⁵ Ubel (1999) assessed the impact of a differences in ability to benefit between severely ill and moderately ill.²⁴ Respondents were asked to allocate resources to severely ill patients that would benefit “a little” from treatment or to moderately ill patients that would benefit “considerably” from treatment. Although the majority of respondents preferred to give an equal amount for resources to both groups (73%), 21%



preferred giving priority to the moderately ill patient group that would benefit considerably from treatment and 6% to the severely ill patient group that would only benefit a little from treatment.²⁴

Bryan et al. (2002) included a criterion in their choice experiment that reflects the chances of success of treatment, in order to test the extent to which people support the QALY maximisation principle for resource allocation.⁵⁹ They found that the strength of people's preference for interventions are proportional to the variations in the chance of treatment success, *ceteris paribus*. The strength of preference was measured by means of the marginal rates of substitution: if intervention 1 has a 50% higher chance of success than intervention 2 and all else is equal, people were willing to sacrifice 50% of the survival gains. This implies that the strength of the preference for the intervention with the highest success is higher if the improvement in the chances of success goes from 10% to 15% than when it goes from 60% to 65%.⁵⁹

4.2.5 Condition-related criteria

4.2.5.1 Severity of disease or medical need

Medical need expresses the *absolute* need for a treatment, regardless of the availability and effectiveness of alternative treatments. Medical need is most often operationalised by using a measure for disease severity. For example, Lim et al. (2012)²⁷ define severity of disease as a combination of quality of life before treatment and life expectancy without treatment.

There are several operational definitions for severity of disease. Severity of disease can be defined as absolute health loss due to illness, in relation to the "fair innings" of people or as a proportional shortfall.⁶⁰ Stolk et al. (2005) found that stated preferences for allocating resources to severely ill are influenced by the operational definition used. They compared three operational definitions and assessed how well each of them correlated with the observed rank of diseases in a convenience sample of students and policymakers. The authors found the strongest support for the fair innings argument. The fair innings argument bases priorities on the number of QALYs foregone, no matter whether health losses occurred in the past or will occur in the future. It starts from the idea that every person is entitled to a number of QALYs in his life with those who have lived a smaller number

of QALYs are worse off and should get priority over those who have lived a larger number of QALYs.

Several studies found a strong preference for giving priority to more severe diseases.^{20, 23, 24, 27, 50, 51} Often the criterion of severity of illness is preferred above the health gains or benefit after treatment.^{20, 23, 24, 27, 43, 56} In other words, people would prefer to give priority to a more severely ill person who benefits less from treatment than to a less severely ill person who would benefit more from treatment. For example, Green et al. (2009) describe that 60% of respondents indicated that a unit of health gain in a severely affected patient group was of greater social value to that same unit of health gain in a moderately affected patient group, all else equal.²³

In a study by Diederich et al. (2012) severity of disease was considered to be the most important criterion for priority setting (weight 50%), followed by current quality of life of patients (as opposed to quality of life without treatment) (weight 24.7%), age (12%), socioeconomic background (7.9%), social responsibility (4.6%) and lifestyle (0.8%).⁵⁰ This study did not consider, however, the improvement in health of a potential intervention, as the objective was to assess the relative importance of patient-related criteria for priority setting according to the general public.

4.2.5.2 Therapeutic need

The criterion therapeutic need is distinct from medical need.² A therapeutic need exists when there is no other treatment available or the alternatives available have limited effectiveness. A *high* therapeutic need exists when the severity of the condition is high, despite the application of current treatment options. Thus, in contrast to medical need that expresses the need in absolute terms, therapeutic need is a *relative* or incremental criterion.

The empirical literature is very often not clear about whether the availability and effectiveness of existing therapeutic options are taken into account when preferences regarding prioritisation criteria are elicited. Disease severity, without further specification, is a very frequently used criterion in choice experiments. Dolan & Tsuchiya (2005), for instance, used a comparator "without treatment" for life expectancy and quality of life, which is reflecting medical need rather than therapeutic need.



We argue that therapeutic need is a more relevant criterion than medical need because a severe illness does not necessarily imply a high therapeutic need. For example, diabetes type 2 is a severe illness, but there are many adequate treatments available already, so the therapeutic need is rather low. Cystic fibrosis, on the other hand, is a severe illness without any treatment and thus with a high therapeutic need. For coverage decisions, therapeutic need is more important than medical need. Focus on medical need would inappropriately lead to ever increasing expenditures for severe diseases for which good alternatives are already available, as innovations in these domains are very often of marginal incremental clinical benefit but more expensive.

The discussion about therapeutic versus medical need is important, because it determines the description of the criteria to be included in an empirical study: should they be described in absolute or in relative terms, compared to an alternative. Our argument is that describing benefits of treatment in absolute terms is misleading.

For example, treatment A may have a higher absolute benefit for a specific severe disease than treatment B for another severe disease, but have a much smaller incremental benefit compared to the existing treatment than B. What matters most is the incremental benefit. Therefore, the criteria should be defined in incremental terms, to allow a relevant judgment of the value of a treatment. Watson⁶¹ uses a comparator for every criterion.

Therapeutic need has yet another dimension. Some studies examined the importance of there being an alternative treatment available or not. If no treatment is available, patients can receive supportive care but no active treatment. These studies thus focus on the fact of there being a therapeutic option regardless of its effectiveness. Some results suggest that people place value on “giving active treatment”.^{21, 51} Linley & Hughes (2013) found that people do prefer treatments for diseases where there are no alternatives available, despite the assumption of little health gain in that patient group compared with considerable improvements in health gain in patients with several treatment options available.⁵¹ Another study found that the availability (or not) of alternatives is as such not an important criterion, compared to the level of health improvement, value for money and severity of disease.⁵⁶

4.2.5.3 Frequency of disease

This criterion is related to the number of people treated or helped by the intervention.

In general, it is assumed that priority is given to more common diseases than to less common diseases. This was the case in a number of empirical studies included in our review.^{21, 32, 51, 61} Especially if the treatment for the common disease offers larger improvements than the treatment for the less common (or even rare) disease, preference is given to the treatment for the common disease.⁵¹

The marginal preference for treating more patients seems to decrease with an increasing number of people already treated. This means that the preference for increasing the number of people treated, will be stronger when the current number of people treated is low than when the current number of people treated is already high. For example, Bryan et al. (2002) show that the QALY-approach, assuming a proportional increase in preference for an expanding programme, underestimated the preference for increasing small programmes, but reflected quite well the preference for increasing large programmes.⁵⁹

Some studies include “rarity of the condition” as a separate criterion.^{32, 51} Linley & Hughes (2012) found that treatments for common diseases that produce considerable improvements in health were strongly preferred to treatments for rare diseases that produce only limited improvement in health.⁵¹ No empirical support is found for the importance of rarity *per se* as a criterion for priority setting. Rather, it is the combination of the severity of the condition, the high therapeutic need and the considerable health improvement after treatment that will *de facto* lead to giving priority to treatments for rare conditions. This clearly illustrates the point we made at the beginning of this review that all findings related to a particular criterion (in this case ‘rarity of the disease’) are conditional upon the other criteria included.



4.2.6 *Intervention-related criteria*

4.2.6.1 *Safety, efficacy and effectiveness of the intervention*

Health benefits in general

The health benefits criterion encompasses the benefits of an intervention on patients' life expectancy and on their health-related quality of life. These benefits can be included as separate criteria or as one criterion encompassing both.^{25, 27, 61}

The current paragraph discusses the results of the studies that considered both types of benefits jointly. The next two paragraphs discuss the results of the studies that used health-related quality of life and life expectancy as separate criteria.

Higher priorities are generally given to interventions with higher health benefits.^{27, 39, 61} Gallego et al. (2007) found that 35% of the respondents considered treatment outcome of the recipients as the most important factor for choosing between potential beneficiaries of high-cost medications.²⁵ Current health status (26%) and quality of life (15%), life expectancy (without treatment) (9.2%), age (9.2%), socioeconomic status (4.6%), family commitments (1%) and lifestyle (0.5%) were all considered less important for setting priorities.

According to the study by Golan et al. (2011), the majority of the people attach greater weight to life-prolongation benefits than to quality of life gains.²¹

Besides the actual value of the health benefit, also the strength of the evidence is important.³⁹ Since medicine is more and more becoming evidence-based one could argue that certain evidence would be needed in order for the treatment to be even considered for reimbursement. Very few studies include this element, although one study found that respondents are more likely to select interventions with a strong evidence base.³⁹ This element was highly statistically significant, besides criteria such as cost, effectiveness of the intervention, and own contribution to illness.³⁹ Yet, the authors warn for a bias effect that people might think that more evidence equals more health gain. Others found that strong evidence on effectiveness has a significant influence on the decisions of respondents, but is not weighted as an important criterion in itself.^{31, 61} Still others, like Johri et al.

(2009), believe that the quality of the evidence is not a relevant criterion to inquire about with the public, but should be left to decision makers.²⁶

In a previous KCE study, it was recommended to reduce the estimated clinical benefit if the uncertainty is large and work with this reduced estimate to determine the relative value of the intervention. Or, alternatively, the decision could be conditional upon further evidence collection (coverage with evidence development). If one of these strategies is followed, there is no need to include the strength of evidence as a separate criterion.² This solution was proposed in the absence of evidence about the impact of uncertainty on the relative importance of decision criteria according to the general public. When following the recommendation, it should be clear that this is not based on citizen's preferences for reimbursement criteria but is rather a pragmatic solution to a problem for which public preferences are unknown.

Quality of life benefits

The studies that investigated the criterion of quality of life show that health-related quality of life is considered to be a very important criterion by the public.^{21, 25, 27, 30-32, 34, 37, 38, 43, 59, 62} Quality of life is often measured together with life expectancy, but as mentioned before when the effects were disentangled, the criterion of life expectancy seems to be more important than quality of life.^{21, 31, 37}

According to a Dutch study 54% of the people agree that quality of life is not stressed enough when deciding upon application of medical technology.⁴⁹ Bryan et al. (2002) point out that the preference for quality of life improvements depends on the health-related quality of life before treatment. The marginal preference declines with an increasing health-related quality of life before intervention.⁵⁹ This means that preferences are much stronger for a change in quality of life from 0,75 to 0,8 than for a change from 0,95 to 1. It should be noted that this study did not look at overall health-related quality of life, but only to health-related quality of life related to usual activities and depression/anxiety.



Impact on life expectancy

The criterion of life expectancy after treatment is a commonly used criterion, that is often found to be very important to the general public.^{21, 25-27, 31-33, 37, 43, 57, 59} Not surprisingly, programmes with larger benefits in terms of life expectancy were given higher weights than programmes with a lower effect on life expectancy.³² Prolonging life is important for many people, even if it is only for a few months. However, the strength of the preferences for survival gains is not constant. Bryan et al. (2002) compared the impact of a life-expectancy improvement of 1 year with a life-expectancy improvement of 5 years.⁵⁹ They found that the preference for a 5-year survival gain is not proportional to the preference of a 1-year survival gain. A possible explanation is time preference, meaning that future life years are considered less valuable than immediate life years. This confirmed the results of an earlier study using the person trade-off technique to assess the relative importance of the duration of benefit on preferences.¹⁴ Nord et al. (1996) found that treating fewer patients with a longer life expectancy was regarded as equally valuable as treating more patients with shorter life expectancy, but that the valuations increased less than proportionately with duration.¹⁴

The National Institute for Health and Care Excellence in the UK (NICE) states that life prolongation becomes even more important in end-of-life situations as long as the life extension is of reasonable quality, but empirical evidence does not support this statement.⁵¹ Confusion might arise when “end-of-life” and “severity of illness” are both included as separate criteria. Severity of illness often contains an element related to remaining life expectancy and thus overlaps with the end-of-life criterion. It is uncertain whether end-of-life would still deserve special attention when it concerns people who are not ill or suffer from a mild illness. As the special status of end-of-life was given by NICE to argue for the coverage of certain cancer medications that would be given at the “end-of-life”, it can be questioned whether this was because it concerned cancer, a severe disease, or whether it was because it concerned an end-of-life treatment. In order to avoid confusion, we would propose to avoid “end-of-life” as a separate criterion next to “severity of illness”.

In cases where a small increase in lifetime outweighs any other benefit, it is referred to as the rule of rescue³⁷.

4.2.6.2 Cost

The criterion of cost for setting priorities in health care is not often investigated. The total costs of an intervention for the healthcare budget are related to the number of patients eligible for the intervention. However, in health economics, not only the number of patients is important, but also the savings induced by the intervention elsewhere in or outside the health care sector are important for the “net” cost of an intervention.

Findings related to costs are only relevant in as far as the opportunity costs of a particular choice are made explicit. These opportunity costs could occur outside the healthcare sector or within the healthcare sector and could be expressed in monetary terms or in terms of benefits foregone. When cost is expressed in monetary terms and studies as a decision criterion for resource allocation within the healthcare sector, a hypothetical budget constraint has to be imposed in the hypothetical choices.

Mortimer (2008)³⁹ found that people make a trade-off between cost, effectiveness and non-health arguments when prioritizing health programmes. Respondents are more likely to select less costly, more effective interventions. Another study contradicts this finding and found that people did not want to give priority to groups of patients who needed less costly treatments, even if this would allow to treat more patients and obtain more health benefits overall, but rather preferred to give equal priority to those who needed more costly and those who needed less costly treatments, all else equal.^{40, 63} Lees et al. (2002)³¹ found that people in the UK prioritize health care that improves health, quality of life or prevents ill health and only later on take cost into account. This means that for the public, the costs are not the major concern when allocating resources.

The question is whether it is realistic to assume that costs should only play a minor role in the prioritization process. Prioritizing health interventions is fundamentally an economic problem; it is because resources are scarce that priorities have to be set. It is unclear whether people are aware of the implications of their choices in a hypothetical choice experiment when they state that cost should not be important as a criterion or, as in Chinitz et al. (2009)³⁸, expensive treatments should be in the top three priorities for resource allocation.



Cost-effectiveness is not treated as a separate criterion in our review, as cost-effectiveness basically combines the criteria of costs and health benefits.^{39, 56} Hence, including cost-effectiveness as a separate criterion would induce double counting, in the sense that the same feature of an intervention (cost and effectiveness) is credited more than once.¹¹ Another issue with including cost-effectiveness as a criterion (instead of cost and effectiveness separately) is that it does not allow to capture differences in preferences for costs versus health effects, as the same cost-effectiveness ratio can be obtained by many combinations of costs and effects.¹¹

4.2.6.3 Innovation

The criterion of innovation refers to the extent to which the new intervention represents a different approach or working mechanism compared to existing approaches or interventions for the same condition; i.e. significant breakthrough or cutting-edge technologies. It does not refer to therapeutic added value. An innovative intervention may have a small added therapeutic value but add to the future prospects of better innovations. In that sense, new interventions are not necessarily considered innovative. For example, a new radio-isotope for Positron Emission Tomography (PET)-scanning that is better able to detect lung cancer in an early stage than the traditionally used radio-isotopes may have an added therapeutic value, but is not an innovation in the meaning of the “innovation” criterion. Electronic nose sniffs⁶⁴ for the detection of lung cancer, on the other hand, are an innovation, because they use a completely different diagnostic approach. Even if these electronic nose sniffs would have no obvious added therapeutic benefit at this moment, they may still be valued high on the list of priorities because they give the prospect of broader applications, in other indications.

In a study by Watson et al. (2012), a distinction was made between cutting-edge technologies and latest technologies.⁶¹ Innovation refers more to cutting-edge than to latest technologies. The study found significant preferences for both types of technologies, suggesting that respondents did not distinguish between latest and cutting-edge technology.⁶¹

Linley et al. (2012) compare the medicine that works in a new way with existing alternatives for the treatment of the same disease. They only found a preference in favour of medicines that work in a new way if they were coupled to a large health improvement.⁵¹

4.2.6.4 Type of intervention

Prevention or treatment

There is some evidence that people value preventive interventions slightly higher than curative interventions.^{25, 30, 39} In the hypothetical scenario where respondents had a limited pool of money they had to spend on two treatments, one preventive, another therapeutic (though not curative), more than 50% of the respondents split the resources evenly between the two treatments, 26% allocated more resources to the preventive intervention than to the therapeutic intervention.²⁵

Another study compared the value of a certain health gain (treatment) to the value of an avoided health loss (prevention).⁶⁵ The results showed that most people (69%) favoured improving health compared to avoiding an deterioration in health, even if the loss is considerably less than the gain. But the preference towards curative treatments was not absolute. Twenty-three percent of participants in the survey favoured cure or favoured avoiding a decline in health. A vast majority of participants was willing to trade their preferences towards cure against a larger number of patients that could be saved from decline.⁶⁵

In a study about the relative importance of prevention, local access, waiting times, national government priorities, and staff time spent with patients for setting priorities, Lees et al. (2002)³¹ found that amongst these criteria, preventive health care gets a high weight from the public.

Care or cure

“People- and family-centred care” and “promoting wellness and strengthening prevention” were considered to be the second most important priorities, after ensuring quality and safety of health care, according to a study by Louviere et al. (2010).⁶²

Lees et al. (2002)³¹ found that people think it is important that the staff spends more time with their patients (more care). Based on the observation that the majority of respondents felt that society should give priority to primary care over technology, Tymstra et al. (1993) concluded that the public prefers to give more weight to care than to cure.⁴⁹



Related to this is the issue of informal care. Linley & Hughes (2013) found that society prefers to give priority to medicines that reduce reliance on informal caregivers, such as family members.⁵¹

Key points

- The literature review focussed on the one hand on general ethical principles for making rational resource allocation decisions in health care on the one hand, and on the other hand on operational intervention-, patient- or condition-related decision criteria.

General ethical principles for resource allocation decisions

- No single general resource allocation principle, be it the lottery principle, the rule of rescue, health maximisation, fair innings or choicism, seems to explain how people want priorities to be set in health care. Almost all empirical evidence finds mixed views. It is probable that variations exist across populations. The choice of the guiding ethical principles is therefore mainly a context-specific choice, depending on the perspectives of citizens about the kind of society they want to live in.
- Empirical evidence suggests that people are prepared to make sacrifices in terms of health maximisation to achieve other societal goals, such as equity and social justice. This means that the traditional cost-effectiveness approach to decision making, with unweighted QALYs, is insufficient to cover all objectives of a health care system.

Condition-related decision criteria

- Severity of disease and health benefits of treatment appear to be two most important criteria amongst all priority setting criteria.

- Therapeutic need, as opposed to medical need, has received relatively little attention in the empirical literature about preferences for reimbursement criteria. Nevertheless, we argue that therapeutic need is a more relevant criterion than medical need, as in many cases an active treatment is already available. Moreover, if “doing something”, irrespective of the effectiveness of that activity, is considered valuable in itself, it is important to clearly make the distinction between therapeutic need and medical need.
- In general, reimbursement for more frequent diseases is preferred to less frequent diseases, but the marginal preference for treating more patients decreases with an increasing number of people already treated. Evidence does not support the claim that “rarity of the disease” is an important separate criterion.

Patient-related decision criteria

- Social worth and socioeconomic status is generally found to be of limited importance in the resource allocation debate.
- Mixed evidence exists also about the relative importance of other decision criteria.
- While many studies find that people prefer giving priority to the young as compared to the old, this preference is not absolute. Age seems to be a proxy for many other things, such as capacity to benefit, expected (quality-adjusted) life years gained, social worth or fair innings.

Intervention-related criteria

- Many studies found that impact on life expectancy is a more important criterion than impact on the quality of life, but the strength of the preference diminishes with an increasing number of life years gained.
- Costs of treatment do not seem to be the most important concern of the public.



4.3 Results: Preference elicitation techniques

Our review of the literature on decision criteria for setting priorities in health care revealed that the results of studies that investigate the relative importance of decision criteria are highly dependent on the preference elicitation technique used and the approach taken for the analysis of the data. The second part of our literature review relates to the techniques for measuring public preferences. A systematic literature review of methods for eliciting public preferences for health care has been published in 2001.¹² We mainly relied on this overview to get an idea of the strengths and weaknesses of different methods. Our review was complemented with more recent research that was not yet included in the review by Ryan et al. (2001).

We only describe the quantitative survey methods, as our objective is to derive quantitative weights for the different decision criteria. A quantitative survey method can be defined as a set of scientific procedures for collecting information and making quantitative inferences about populations.⁶⁶

Ryan et al. (2001) classified the methods for preference elicitation in three groups: ranking, rating or choice-based techniques.¹² Six studies included in our review used a ranking method, five a rating method and twenty-three studies a choice-based method (see tables in appendix).

Before discussing the validity, reproducibility, internal consistency and acceptability of different preference elicitation techniques, we would like to draw attention to the underlying implicit or explicit assumptions about human behaviour in each technique (4.3.1). This is important for the later interpretation of the survey results.

4.3.1 Assumptions about human behaviour

It is hard to find out what the true preferences of people are, especially in the field of health care priority setting. Researchers have created different methods based on different assumptions of human behaviour. Transparency about these assumptions is important. Nevertheless, most research about health care priority setting is not explicit about them.

First, we describe the assumptions about people's interpretation of criteria. Then, we describe possible epistemological assumptions about how people process information, and finally we describe assumptions about how people interact with question framing. An overview of the strengths, weaknesses,

threats and opportunities associated with each of these assumptions is provided in Table 10.

4.3.1.1 Assumptions about the interpretation of criteria

When criteria are selected at the start of a project, assumptions are made about the interpretation of these criteria. Many researchers do not provide an explicit definition of the criteria under scrutiny. Instead, they rely on a universal intuitive connotation or definition of therapeutic need, age, life expectancy, quality of life, responsibility etc.

There is also a possible effect of the question framing. Depending on which criteria are presented in the same exercise, people may give a different meaning to the same criterion. For example, people may give a different meaning to the criterion quality of life when life expectancy is also included than when life expectancy is not included. Hence, techniques that ask to consider criteria one by one might give different results from techniques that ask to consider a set of criteria at once.

4.3.1.2 Assumptions about the functional specification of preferences

It is usually assumed that people's preferences with respect to criteria are linear and independent. However, both of these assumptions might be too strong. Weights may change depending on the baseline level of a criterion (i.e. the current situation). By using on single weight for the criterion "quality of life improvement", for instance, irrespective of the baseline level of quality of life, we ignore the fact that patients might assign different weights to this criterion if quality of life is already high than if quality of life is very low.

It has also been observed that preferences for criteria are often conditional upon the value of other criteria, i.e. weights of attributes might change if the level of another attribute changes and thus preferences are not linear. It implies that the rate at which patients trade-off different attributes varies according to the levels of each attribute (in economics, this is called the "marginal rate of substitution between attributes").

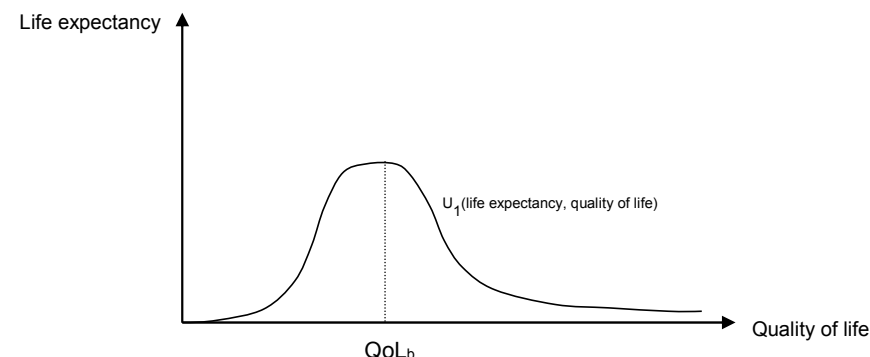


For example, suppose people only consider current quality of life and life expectancy as relevant attributes determining the level of therapeutic need. The rate at which people wish to trade-off quality of life (attribute 1) against life expectancy (attribute 2) might depend on the level of quality of life and life expectancy. For instance, people may be reluctant to give up quality of life for one year of life extra if current quality of life is low, while they may be willing to trade-off quality of life for a longer life expectancy if current quality of life is very high. Thus the willingness to trade-off is low when quality of life is low and is relatively higher when quality of life is higher.

The willingness to trade may even become negative; i.e. at very low levels of quality of life, people might prefer to live less long or to get a better quality of life if life expectancy would increase. This is graphically shown in Figure 6. All quality of life-life expectancy combinations on the curve are equally valuable for the individual, or, they are all points on the same utility curve $U_1(\text{life expectancy, quality of life})$. The convex nature of the utility curve from QoL_b onwards, shows the diminishing marginal rate of substitution if quality of life increases. Values for quality of life below QoL_b are by this person considered so bad that he/she would rather live shorter in this health state than longer.

The consequence for MCDA, where criteria are weighted by their relative importance, would be that there is no single constant weight for each of the criteria but rather a range of weights that apply depending on the baseline level of the criterion *and* depending on the level of other criteria included in the MCDA. Translated to our example: the weight for quality of life improvements would be higher if current quality of life is low than when current quality of life is already high; and the weight for quality of life improvements could be higher, the shorter the remaining life expectancy.

Figure 6 – Changing marginal rate of substitution between attributes



Some techniques do not allow for identifying changing marginal rates of substitution or interactions between criteria. If it is assumed that preferences for one criterion are conditional upon other criteria or preferences for one criterion interact with preferences for another, the chosen method and analysis technique should allow for the identification of such relationships.

Another issue related to the specification of preferences is overlap between criteria. For example, the criterion age may overlap with the criterion expected therapeutic benefit, as people expect younger patients to have a higher therapeutic benefit than older patients. It may imply that in a direct comparison between two criteria, one is falsely preferred over the other because it overlaps very strongly with a much more important criterion. For example, in a direct comparison, age could be considered more important than socioeconomic status. However, this can be due to the fact that therapeutic benefit – of which age is perceived as a proxy – is considered much more important than socioeconomic status, rather than age as such.



4.3.1.3 *Meta-ethical assumptions*

Meta-ethical assumptions are assumptions about how people think about ethical dilemmas in health care priority setting.⁶⁷ Some methods are based on the hypothesis that ethical thinking in health care priority setting follows the principles of rational thinking, implying consistency. This paradigm of rational ethical thinking can be called **cognitivism**.⁶⁸ Others assume that ethical thinking in health care resource allocation is rather emotion-driven. This paradigm of emotional ethical thinking is called **emotivism**.⁶⁸ Yet another position holds the **middle ground** between cognitivism and emotivism. Depending on the assumptions one holds about ethical thinking, one could construct the cases in a way that stimulates a certain way of thinking. Sometimes consistency tests are used to test the meta-ethical assumptions.

Utilitarianism falls under cognitivism.⁶⁹ When it is believed that people think in a utilitarian way, it is assumed that people will also respond in a consistent utilitarian manner.

A strict rational methodology presumes that people answer consistently in a way that maximises their utility. A consistency test or transitivity test to see if people answered in a rational way is often used. If people do not choose in a consistent manner, the researchers assume that the respondents did not respond according to their real preferences which are rationally consistent. Approaches that disregard inconsistent results could result in preference sets people would ultimately not support, but consistency is indispensable in a normative approach according to choice theory in economics (see paragraph 3.1.4.).

An emotive methodology lets people make choices purely based on their emotions. It often also allows respondents to make qualitative statements to explain their feelings about the question or allowing a “don’t want to answer” response. However, this could induce strategic behaviour, in that people use a “don’t want to answer” option to escape the dilemma. “Don’t know”-answers may be an indication of incomplete preference orderings, as described in paragraph 3.1.4. Fully emotional and incomplete responses are not necessarily inconsistent or –in that sense- irrational. From a policy point of view, the results of such studies may be informative, but less helpful because at the end of the day decision makers have to make a decision. As long as the information they get from the public, based on this methodology,

is “we don’t want to choose for this or that reason” or is purely based on the emotion of respondents, it is hard to decide which weight should be given to these preferences in a societal decision context. Emotive methodologies are appropriate, however, for explaining observed choice behaviour and could, as such, be a useful complement to the quantitative rational approaches discussed before.

4.3.1.4 *Epistemological assumptions*

Next to assumptions about moral thinking, some assumptions are made about the amount of information people can process to make a decision about health care priority setting. These are called epistemological assumptions.⁷⁰ A certain amount of information would make questions too complicated and people would not understand the question or would be unable to answer. In the end the researcher wants to obtain weights for different criteria that are under scrutiny. Some information is related to criteria, other information is given as a context.

Most researchers assume that the respondents cannot take too many criteria and principles into account at once when making a decision. An option is to reduce the number of criteria per question and let respondents consider only a few criteria spread over several questions. Following this methodology, the criteria are first divided over several questions and subsequently a general conclusion is made about the weights based on the separate smaller weighting exercises.

The approach of little information can be used with any method. For example, a ranking exercise can be done with a limited number of criteria, Likert scale questions can be asked about a few separate criteria, choice experiments with only three criteria, etc. The analytic hierarchy process is a method that is specifically constructed so that respondents have to express preferences about bits of elements. The criteria are first weighted in a pairwise manner: two criteria are directly compared with each other and a weight is assigned to each of them relative to the other. Afterwards, concrete cases (e.g. two anti-depressants) are compared on each of the criteria and weighted. Some techniques, such as the “best-worst scaling technique” (see paragraph 4.3.2.3), assume that it is better to allow the respondent to choose among a limited set of options and express the best option and the worst option.



Assuming that people need all criteria to judge upon a situation to make a balanced decision would typically lead to a method which would present multiple criteria at once. Examples include ranking exercises with multiple criteria, Likert scale questions with multiple criteria and discrete choice experiments.

In addition to the information about the criteria themselves, less or more contextual information can be given, or contextual information could be provided only if asked for by the respondents. Too much information might induce escape behaviour because the task becomes too complex or emotional responses if the respondent happens to be in the case described (e.g. when vignettes are used). Too little information, on the other hand, might not give valid answers because people lose sight of other ethical principles, opportunity costs or information about the budget impact.

Some people do not understand the question and perhaps some people who understand the question do not support the premises (i.e. a choice becomes irrelevant when the criteria are irrelevant). Depending on the assumptions about the potential of people to understand the questions, more or less response options could be provided. To avoid people not answering because of complexity or emotional reasons, some researchers made other options available next to choice A or B as “undecided” or “don’t know” or “don’t want to answer”.

4.3.1.5 Behavioural assumptions: the interaction with the question framing and with the data collection tool

Other important aspects of research on societal preferences for resource allocation in health care are the experimental conditions (i.e. personal contact with an interviewer, web-based survey, anonymous or not, etc.) and how people react on the framing of questions. Some researchers assume people will never state their true preference in an experiment, because people never choose in a hypothetical experimental situation the same as they would in real life. Some framings try to come therefore as close to reality as possible, by presenting the respondent real-life cases. However, most people are not likely to have to make choices between patients/treatments in real life people. Thus, the question is whether it is really important for a normative analysis to stay as close to real cases as possible.

The extent of personal recognition also plays a role. Some methods ask about the patients themselves, their close relatives or patients and health conditions they know very well. Others consider this approach to be too subject to bias and unsuitable for measuring general preferences for priority setting, because the responses would be biased by self-interest. Ideally, the questions in the survey should find a balance to stimulate a personal attachment to the case that would not be too big so it would avoid focusing too much on personal interests that would overrule societal interests.

Some discrete choice experiments, where people are asked to choose between two or more scenarios in order to derive their preferences with respect to specific scenario features (see paragraph 4.3.2.3), describe scenarios in abstract terms, while others describe scenarios based on real-life cases, with extensive descriptions. The latter may trigger intense emotions. Which information should and which information should not be provided and how the information should be provided is a tricky methodological question for the kind of study we envisage. It relates also to epistemological assumptions about the amount of information people can process and the meta-ethical assumptions about the risk of bias.

One study tested the influence of giving more or less clinical information to the respondents on their results.⁷¹ It was found that people indeed tend to choose differently if more clinical information is provided



Table 10 – Overview of strengths, weaknesses, opportunities and threats of different behavioural assumptions

	Options	Strength	Weakness	Opportunity	Threats
Assumptions about moral thinking of people	Universalism: the meaning of ethical criteria is roughly the same for everyone	Easy to draw generic conclusions	No specific information is provided about the criteria under investigation	A moral reasoning exercise can be given before the experiment to stimulate the moral thinking	People use a different meaning for the same criterion in different exercises, which leads to results that are difficult to interpret
	Relativism: the meaning of criteria varies between individuals ⁷²	More realistic	Difficult to operationalise, difficult to draw generic conclusions from results		
Meta-ethical assumptions: the way people think about criteria	Cognitivism: rational thinking: people have complete preferences (reflexive, consistent and transitive)	Fits in the economic utilitarian paradigm. Easy for quantitative data-analysis	No place for emotional thinking or an emotional response	Test with a consistency or transitivity test	Because there is no space for emotional responses and inconsistent results are ignored the results could lack validity
	Emotivism: people think in an emotional way	Cases can contain concrete emotional personal elements	Inconsistent results are more difficult to handle in aggregated analyses	Option to answer: “don’t want to answer”, and option to state emotions	People could use “don’t want to answer” as an escape route
Epistemological assumptions: the manner in which people process information	Providing a lot of information	Many criteria are evaluated at the same time, respondents are free to choose what to take into account	Too complicated for the respondents	A method might cope with this by allowing a “don’t know” response	Too much info, might induce strategic behaviour
	Providing little information	Simple acceptable exercises for respondents, paternalistic approach	Loss of complex reality		Too little information, might not give useful responses



Behavioural assumptions: how people interact with question framing and experimental conditions	Realistic concrete (labelled) cases	Cases have a more realistic set-up	Cases might be more complex to understand and stimulate responses inspired by self-interest	Data collection tools as face-to-face interviewing might make the cases more realistic	People might choose differently if specific clinical information is provided, because they personalize too much and lose sight of the concept of general priority-setting
	Abstract (non-labelled) cases	Cases are easier to construct and analyse for the researcher. Allows respondents better to disconnect from real-life examples or experiences. Could simulate a general priority setting context	A more paternalistic view that provides abstract cases would not allow the respondents to answer as they would in real life		Insufficient affiliation with the sensitive matter of health care priority setting

Key points

- The measurement of preferences regarding reimbursement decision criteria requires simplified assumptions about human behaviour, including how people process information, how they think about the criteria, which implicit assumptions they make and whether or not preferences for specific criteria are independent of the value of other criteria.
- Preferences for reimbursement criteria are inevitably in part influenced by emotions. This might be a problem, as emotions are not stable. To derive useful preferences for policy making it is important to use techniques that avoid emotional responses.
- Any preference elicitation approach should avoid information overload or cognitive impossible valuation tasks. A good balance between amount of information and feasibility should be pursued. More information can be more precise but less feasible, too little information might reduce relevance.

4.3.2 Techniques for measuring preferences

4.3.2.1 Ranking techniques

Ranking techniques ask respondents to give a ranking of priorities for health interventions. They allow the respondent to rank his preferences in a comprehensive list ranking for instance from 1 to 10. The options that receive the highest ranking are viewed as the most important.

In ranking exercises, higher rankings are often found for children and people who are dying.^{18, 30, 34, 43} However Dolan & Tsuchiya⁴³ indicate that the results depend on the age ranges used. Studies who compared the ranking exercise with a rating method (Likert scale) found similar results regarding the preference to give priority to the young.^{30, 34.}

4.3.2.2 Rating techniques

Rating techniques ask respondents to express their opinion about a set of scenarios or statements on a numerical or semantic scale. Semantic scales, such as the Likert scale, can range from “I completely disagree” to “I



completely agree” and typically have a neutral point in the middle. A numerical scale, such as the Visual Analogue Scale (VAS), uses two extreme anchors i.e. between -5 and +5 and asks respondents to indicate the point on that scale corresponding with their preference. The outcomes of a rating technique can be used to constitute weights.¹²

The most important advantage of rating techniques is that they measure the strength of preferences. Scales can be discrete or continuous. The number of intervals used in case of a discrete scale or the length of the scale in case of a continuous scale could influence the responses. A long scale allows theoretically for finer preference indications. However, compared to the intervals on a five point Likert scale, a long scale also implies smaller intervals between answers as, because in principle an infinite number of intervals are possible. It makes the interpretation of the results rather complex. However, also within the Likert scale people can assume an unequal interval between, for instance, agree and strongly agree and between agree and undecided.¹² Such implicit differences in intervals, implying differences in strength of preference, are unknown to the researcher analysing the data.

Baron and Ubel⁷³ found with the VAS exercise that mild conditions are often rated quite high in terms of priority compared to other conditions. The results were similar to the results of a person trade-off method. After the exercise with a VAS scale,⁷³ the respondents were confronted with their priority ranking. They often adapted the ranking and gave more weight to those with a larger health gain from the treatment. It can be concluded that rating alone does not always give a good overview of respondents' actual preferences for priority setting. The combination of rating and ranking may be a solution to get a more accurate idea of people's preferences.⁷³

4.3.2.3 Conjoint analysis and choice-based techniques

Conjoint analysis is a method used to elicit the relative importance of different attributes (or criteria in the current study) of goods and services for consumers. Conjoint analysis is based on random utility theory. It assumes that people derive utility from the consumption of a good because of the characteristics of that good. Hence, the utility function of an individual is determined by the characteristics of the goods and services consumed. Similarly, the characteristics of health care programmes determine the utility of these programmes. By means of conjoint analysis, researchers try to

derive the relative value or utility of health care programmes. Conjoint analysis can be performed using ranking, rating, or choice based techniques. In the context of preference elicitation regarding health care priority setting, it is mainly used in combination with choice experiments.

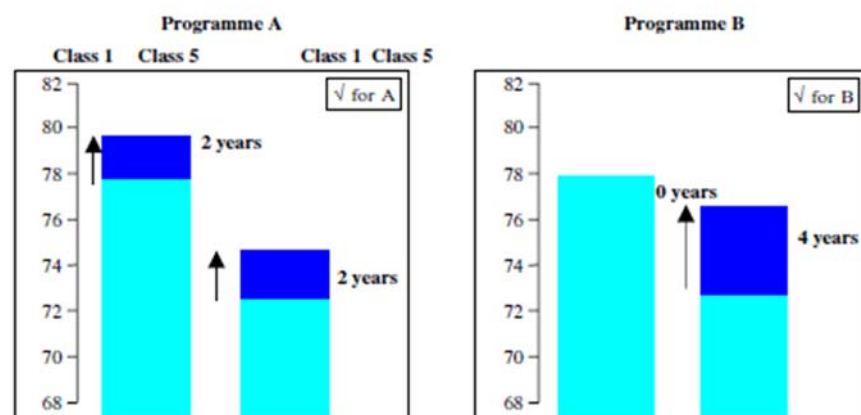
The simplest form of a choice based method presents two scenarios that vary with respect to one or very few characteristics, and ask the respondent to make a choice.

Tsuchiya et al. (2007)⁵⁷ used a choice method to find out about the relative weight of socioeconomic characteristics and life expectancy, given that people are informed about the fact that life expectancy differs at baseline between socioeconomic classes. Figure 5 shows an example of a choice question, comparing two programmes with different outcomes in terms of life expectancy for people in the same socioeconomic class.



Figure 7 – Example of a choice question

As you might know, average life expectancy differs by social class. There are differences between people in social class 1 (for example, doctors and lawyers) and people in social class 5 (for example, road-sweepers and cleaners). These two groups are more or less equal in size (they each make up about 7% of the population). Whilst actual life expectancy varies between individuals, on average, people in social class 1 live to be 78 and in social class 5 they live to be 73. Imagine that you are asked to choose between two programmes which will increase average life expectancy. Both programmes cost the same. In the two graphs below, the light grey part shows average life expectancy, and the dark grey part shows the increase in life expectancy. There is a separate graph for each of the programmes. As you can see, Programme A is aimed at both social classes and Programme B is aimed only at social class 5. Please indicate whether you would choose A or B by ticking one box.



Source: Tsuchiya (2007)⁵⁷

More complex choice methods construct scenarios with multiple criteria (or attributes). A typical form of this method is the **discrete choice experiment**. First, criteria are selected that are considered important by the public for health care resource allocation. We call these “attributes”. Next, levels are assigned to each attribute. This results in a large number of potential scenarios (the more attributes and the more levels per attribute, the higher the potential number of scenarios). In order to reduce the number of choices to be made, factorial designs are used. Respondents are presented a limited number of questions involving a choice between two scenarios. Scenarios differ on several criteria at once. After the exercise, the data are analysed by using regression techniques. The relative importance or weight of the separate criteria, the rate at which individuals trade between these criteria and the overall preference for different combinations of levels of criteria, can then be determined.¹²

Q3. Would you prefer the government to implement 3A or 3B? (Pair 29)

	KEY FEATURES
3A	<p>↓</p> <p>A medical program to prevent a health problem from occurring in working-age adults. The problem is not caused by patients' behaviour. Based on strong evidence, the program is expected to save 40 lives every year. It will cost ten million dollars. Patients will pay half of the cost of their participation.</p>
3B	<p>A lifestyle program to prevent a health problem from occurring in young adults. The problem is partly caused by patients' behaviour. Based on strong evidence, the program is expected to save 20 lives every year. It will cost one million dollars. Patients will pay half of the cost of their participation.</p>

Tick ONE box ☒ to indicate which program you prefer:

☐ 3A
 ☐ 3B

Briefly, what are your reasons for this decision?

.....

.....

Another choice based method is a **person trade-off**. A person trade-off (PTO) asks the question who to treat and assumes that both severity of health and a potential for health after treatment are important criteria. Respondents need to indicate which group they would prefer to treat if there are x people in an adverse health situation A and y people in an adverse health situation B. x or y are varied until the respondent is indifferent between the two options. The social value of a health intervention for B would then be equal to x/y because B is x/y as undesirable as A. This approach has been challenged for its underlying assumptions, such as the transitivity principle, implying that if people consider treatment of 1 patient in state A to be equivalent to treating 10 patients in state B, and 1 patient in state B to be equivalent to 10 in state C, then they should find 1 in state A equivalent to 100 in state C.^{74, 75} Baron et al. (2001)⁷³ used the following question “How many people saved from death is just as attractive as providing 100 people with [the condition-treatment pair].”⁷³ Other studies

The **pie-method** (allocation of points) often uses a visual representation of the total budget (circle) and asks respondents to cut the pie into pieces in correspondence with their preferences of health care resource allocation.¹² A variation of this technique has been used by Wiseman et al. (2005)¹⁸ Respondents were asked to allocate \$A10 million to different health programmes, treatments or groups of the populations (see Figure 9). The advantage of the pie method is that it makes the opportunity costs explicit. If more resources are devoted to one intervention, less remains available for the other interventions. Some also use a technique that does not ask to allocate a budget but where in the total package the budget should be diminished or enlarged.^{76, 77}

**Figure 9 – Example of a question using pie method**

3. If you answered 'Yes' to question 1a please indicate below how you would spend an extra \$A10 million across these 3 programmes? (Remember the total amount must equal \$A10 million). Additional information about the cost and the effectiveness of these programmes can be found on the attachment to this questionnaire.

Parents of young children with behaviour problems This is an intensive training programme aimed at parents of young children aged 3 to 8 years with behaviour problems	= \$ _____
Influenza vaccine This vaccine is aimed at helping to protect vulnerable groups (such as the elderly) against catching the influenza virus that can cause pneumonia and respiratory illness	= \$ _____
Smoking reduction programme In this programme children approaching adolescence receive education and practical help at school to avoid taking up smoking and to quit smoking	= \$ _____
TOTAL	\$A10 million

Source: Wiseman et al. (2004)¹⁸

The **best-worst scaling technique** is based on a theory about how people make best and worst choices from choice sets consisting of three or more elements. The elements can be anything, from criteria to complete interventions or diseases. However, in this context we are mainly interested in choices about decision criteria. The goal of a best-worst scaling choice

experiment is to rank criteria and calculate their relative distance on a utility scale (see Figure 10 for a general example). This is repeated until multiple combinations have been presented to the respondents. As with discrete choice experiments, the number of scenarios to present to respondents can be reduced by using factorial design. Based on the results of the individual trade-offs, the weight of every criterion can be calculated.⁶² Variants of best-worst choice experiments are the “best-worst scaling case 2” and the “best-worst scaling case 3”. The difference lies in how the criteria are judged to be “best” or “worst”. In the case 2 version, respondents have to choose the best and worst criterion out of a list (see for instance Figure 10). In a best-worst scaling case 3, several criteria are combined into three scenario descriptions, amongst which the respondent chooses the best and the worst scenario (see for instance Figure 11). This case actually boils down to a discrete choice experiment with three alternatives, in which respondents make two choices, one about the most preferred alternative, one about the least preferred alternative (cfr infra).










Figure 10 – Example of a general best-worst scaling exercise – case 2

Most Important (Tick ONE box)	Of these, which are the most and least important?	Least Important (Tick ONE box)
<input type="radio"/>	Being self-fulfilled in life	<input type="radio"/>
<input type="radio"/>	Having security in life	<input type="radio"/>
<input type="radio"/>	Having warm relationships with others	<input type="radio"/>
<input type="radio"/>	Having fun and enjoyment in life	<input type="radio"/>
<input type="radio"/>	Having an exciting life	<input type="radio"/>
<input type="radio"/>	Obtaining a sense of accomplishment in life	<input type="radio"/>

Source: Lee (2007)⁷⁸



Figure 11 – Example of a best-worst scaling exercise – case 3

	Treatment 1	Treatment 2	Treatment 3
	I receive medication via tablets	I receive medication via a pump	I have to undergo brain surgery
	I seldom to never suffer from tremors.	I often suffer from tremors.	I seldom to never suffer from tremors.
	I seldom to never suffer from posture and balance problems.	I often suffer from posture and balance problems.	I sometimes suffer from posture and balance problems.
	I seldom to never suffer from slowness in motion.	I sometimes suffer from slowness in motion.	I seldom to never suffer from slowness in motion.
	I seldom to never suffer from dizziness.	I sometimes suffer from dizziness.	I often suffer from dizziness.
	I sometimes suffer from drowsiness.	I often suffer from drowsiness.	I sometimes suffer from drowsiness.
	I often suffer from rapid uncontrolled movement.	I sometimes suffer from rapid uncontrolled movement.	I seldom to never suffer from rapid uncontrolled movement.
Which treatment do you find most desirable?	1 	2 <input type="checkbox"/>	3 <input type="checkbox"/>
Which treatment do you find least desirable?	1 <input type="checkbox"/>	2 	3 <input type="checkbox"/>

Source: Janine van Til (personal communication)



Analytic hierarchy process (AHP) is a method that was introduced to support strategic decisions in the industry.⁷⁹ It is an approach where a multi-attribute decision problem is structured into a hierarchy of interrelated elements (see Figure 12). Independence across criteria is assumed. Analytic hierarchy processes have been applied within health care research to support shared decision making between patient and doctor about certain treatments.^{79, 80}

AHP uses a rating scale to express preferences for criteria and to express preferences for alternatives. Typically, AHP consists of two steps. First, each relevant criterion is compared with each other relevant criterion in a paired comparison, as in Figure 12. People determine the relative importance of one criterion compared to another by selecting a number on the scale. The closer to the left hand side, the more important the respondent considered life expectancy relative to quality of life. If the left hand side 9 is selected, the respondent considers life expectancy to be the only important criterion and quality of life to be not important. The same exercise is repeated for every possible pairwise comparison. This forms a comparison matrix with calculated weights, ranked eigenvalues (representing the weights), and consistency measures.⁷⁹ Then, respondents are asked to rate each alternative (e.g. concrete disease) on each of the criteria, independent of the relative importance attached to the criteria. The eventual choice between alternatives can then be made by combining the scores and the weights for a specific case as a weighted average score, as in MCDA.

Figure 12 – Phase 1 of the analytical hierarchy process: paired comparison between criteria

In order to determine the need of patients for a better treatment, which feature of the patients' life under current treatment is more important to you and to what extent?	
9 – 8 – 7 – 6 – 5 – 4 – 3 – 2 – 1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9	
(his/her life expectancy)	(his/her quality of life)
9 – 8 – 7 – 6 – 5 – 4 – 3 – 2 – 1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9	
(comfort of treatment)	(his/her quality of life)
9 – 8 – 7 – 6 – 5 – 4 – 3 – 2 – 1 – 2 – 3 – 4 – 5 – 6 – 7 – 8 – 9	
(his/her life expectancy)	(comfort of treatment)

4.3.3 Evaluation of preference elicitation techniques

As described by Ryan et al. (2001)¹² there are four criteria to evaluate the methods for preference elicitation: validity, acceptability, consistency and reproducibility.

4.3.3.1 Ranking techniques

The acceptability of ranking exercises is high. Ranking has proven to be relatively easy to complete and easy to analyse.¹² In combination with a face-to-face interview technique, response rates on ranking techniques varied between 75%³⁰ and 83%¹⁸. The response rate of a postal questionnaire with a ranking exercise is usually much lower; e.g. 13.2 % in Dolan et al. (2005)⁴³

The evidence on the internal consistency is relatively weak. Mak et al. (2011)³⁴ and Bowling et al. (1996)³⁰ found that people rank young people very high and older people very low, suggesting a preference for treatment of younger people over the old. However, the same people gave in a rating exercise a very high weight to "everyone should have treatment", suggesting that the old should be treated in the same way as the young. This finding cannot prove internal inconsistency of the ranking technique, however, as the disparity in stated preferences could have been influenced by the wording and interpretation of the questions.¹²

No data are available about the reproducibility of ranking exercises.

Ranking techniques have limitations. First, because the decision-making context is most often lacking, respondents will not take the potential consequences (opportunity costs) of their choices into account when ranking options. Second, due to a lack of information about the strength of preference for each criterion, it is difficult to judge how important each criterion is in the list. One can at best only conclude that some criteria are more important than others. Third, when analysing the results of a ranking exercise, trade-offs between different criteria are not visible. Finally, since no options can be left out of the ranking or be put on an equal level in a ranking exercise, the results may not be valid. In some studies there are a lot of respondents that prefer equal preference, so leaving these out reduces the validity of the results.¹⁸



4.3.3.2 Rating techniques

Likert scales seem relatively easy to complete.¹² Some claim that rating scales are easier to answer than ranking scales.⁷⁸ Response rates as high as 95% have been achieved⁴⁹; 71% by telephone²⁹ and 15%-51% by post.^{28, 31}

Another observation with respect to rating scales is that different people may use different parts of the scale and some response options (e.g. the extremes) are systematically avoided by some respondents. This affects the mode, median, mean and variance obtained.⁷⁸ Some claim that an important disadvantage of rating scales is precisely that people do not use rating scales consistently.⁶²

No information about reproducibility was found for rating studies.

4.3.3.3 Choice-based techniques

Choice-based methods require the respondents to choose between two or more alternatives. The act of choosing is believed to trigger preferences of respondents that lie close to reality. Some methods ask explicitly to choose between comprehensive cases, others decompose a decision problem into little choices and recompose these into overall preferences. When respondents are asked to express their preference on pairwise presented criteria or scenarios, one could argue that respondents lack the impact of their stated preferences on the general ranking of scenarios. This is the trade-off that one has to make when analysing the choice process in this manner. How good can someone express his or her opinion on a fragment of the decisions in comparison to a decision that takes into account all information at once? A possible solution could be to use a computer-based data collection system and order the programme to give feedback along the exercise. This system could provide an up to date score of the overall weighting of the criteria every time a question is completed.

In choice experiments regarding health care priorities people might not want to answer. Making choices between groups of patients is often found to be morally difficult. Some studies excluded non-answers, others performed a sensitivity analysis of the answers⁵⁴ and still others allowed people to state the reason why they did not (want to) choose.³²

In order to improve the quality of responses in choice methods, Johri et al. (2009)²⁶ carefully evaluated the effect of giving one group a moral reasoning exercise and the other group no exercise. They found that the intervention group, i.e. those respondents performing a moral exercise before the choice experiment, gave more “no preference” answers and less extreme answers favouring the young than the control group. Moreover, people from the intervention group did not put as much weight on the lifestyle factor in a coronary bypass scenario as the control group. This might indicate that if people perform a moral reasoning exercise before they complete the choice exercises, they tend to have a more balanced idea of criteria or are more doubtful about extremes especially related to age and lifestyle.

Simple choice methods

Regarding consistency, Johri et al. (2009) found good consistency between answers given to simple choice questions where only age differed between scenarios.²⁶ Consistency was tested by presenting two identical choice sets to respondents, with only the tag being different between the choice sets: the first set concerned a choice between two patients needing a liver transplant, the second between two patients needing a lung transplant; only the age of patients was different. Responses were considered consistent if respondents made the same choice in both choice sets; i.e. twice either for the young or the old patient. The authors also assessed temporal stability of preferences. They found that people who were randomised to the group who did not do a moral reasoning exercise before answering the choice questions had less stable answers regarding age than people who were randomised to the group that performed a moral reasoning exercise.

Discrete choice experiments

Seven out of 23 papers using choice-based methods included in our literature review, used a discrete choice experiment (DCE) to elicit preferences for health care priority setting criteria. Discrete choice methods are based on random utility theory. In the statistical model used to analyse DCE data, it is often assumed that people have continuous preferences, meaning that there is always one criterion or one level that can be traded against another when making a choice who should receive priority for care. Empirical studies found, however, that this might not be the case in real life for some people, as demonstrated by the so-called non-traders. There is still discussion about the interpretation of these people's responses in the



framework of this methodology.¹² A way to find out whether people are actual non-traders or just uncertain about their choice (but still make a choice), is to add a “don’t know” and a “don’t want to answer” response option.

Discrete choice experiments are sometimes experienced as difficult by respondents. It has been suggested that the amount of choices should not exceed 12 per individual. Moreover, in order not to make the choices too complex the number of criteria should not be higher than 5 to 6 and the levels should be chosen in function of a reasonable complexity.¹²

In order to deal with the complexity of 10 attributes, Watson et al. (2012)⁶¹ spent 45 minutes before the actual survey, explaining what each attribute meant and how the results would be used. In this way, they tried to anticipate to the critique on discrete choice methods that they induce strategic choice behaviour. However, some argue that considering several attributes jointly and forcing the respondent to make a trade-off would result in choices that are less related to self-interest and therefore closer to what they would like to pursue for society.⁵⁰

Roberts et al. (1999)⁷¹ explicitly provided a different amount of clinical information about the patients in three respondents groups and found that the more information was given to the respondents the less preference was given for a resources allocation according to the QALY maximization principle. This could suggest that people become more emotional or triggered in the way that efficiency is less important if more clinical information is given. Also Smith et al. (2003) emphasized that the challenge is to provide clearly defined information about attributes to avoid that respondents fill the “information gaps” with default assumptions concerning scenario characteristics.⁸¹

For the analysis of the results of a discrete choice experiment complex statistical models are needed. These models are based on assumptions. How exactly the criteria are believed to interact with each other and how they should be interpreted are methodological assumptions that can change the results. However, the underlying assumptions of the data analysis are often not explicitly mentioned in papers reporting on empirical studies. Some

assume a linear relationship between attribute levels and calculate the weight for each attribute, others assume that there is no linear increase in benefit within an attribute and calculate weights for each attribute level (e.g. Watson et al. (2012)⁶¹).

Discrete choice experiments in combination with a computer assisted face-to-face interview gave a response rate of 94.3% in one study.⁵⁰ Others also find a high participation rate if respondents are first invited by post and take part in a face-to-face interview afterwards: 86%-99%.^{61, 71} Discrete choice experiment questionnaires sent by post generally have a much lower response rate. In Mortimer et al. (2008), the response rate was about 16%.³⁹

As for the acceptability of the discrete choice exercise in itself, the proportion of “don’t know” and “refuses to participate” are a good indication. Some studies only had 2% “don’t know” answers on every question⁵⁶, others received up to 26.2% no responses and 22.2% refusals to participate.⁵⁹

Consistency tests, involving tests whether a dominant alternative^a is preferred to a dominated alternative, show that between 3% and 6% of respondents fail this test.^{39, 56, 71}

An empirical study by Bryan et al. (2000)⁸² found good reproducibility of discrete choice experiments (test-retest kappa statistic of 0.71 for responses to identical choice questions on the same day and 0.65 for responses after 2 weeks’ time).⁸² Moreover, the study found that random effects probit models based on the test data and re-test data were not statistically significantly different, giving support for the good test-retest reliability of the conjoint analysis.

Pie method

The pie method is often praised because it gives a realistic sense of the opportunity costs policymakers face when they have to make resource allocation choices.

Regarding the amount of contextual information, Wiseman et al. (2004)¹⁸ did not find an effect on the outcome of a pie method when giving more information about the costs or expected outcomes of a health care programme.

^a A dominant alternative is an alternative that is better on all attributes than the dominated alternative.



The acceptability for pie methods is as good as 83% when used in face-to-face interviews¹⁸ and 84% when used in combination with a web survey³². Using a postal questionnaire, Kinnunen et al. (1998) found a response rate of 59%.⁷⁶ Another study found a response rate of only 65.1% for the pie method administered during a face-to-face interview.⁷⁷

Schwappach et al. (2006)³² found that 95% passed the dominance test with a budget pie method: people passed when they allocated a higher amount of points (proxy for budget) to the one scenario out of two that was better on all attributes. Seventy-five percent also passed the transitivity test, which tracks if someone prefers A over B and B over C he/she also prefers A over C.

When comparing the results over time and asking the same sample of people to retake the pie method exercise after 34 days, the authors noticed that respondents tended to discriminate more strongly between health programmes, e.g. by allocating significantly more or less points to interventions.³²

Analytic hierarchy process

The greatest challenge for the validity of the AHP method is the ranked reversal phenomenon. Rank reversal happens when the introduction of another criterion induces a change of the desirability of a specific criterion. This happens if the criterion under consideration is as such not considered very important but it is closely related to another criterion that is highly important.⁸⁰ However, new adjustments of the AHP method should be able to preclude rank reversal.¹²

Rank reversal could also happen in other techniques, but is less likely to occur in techniques that ask respondents to consider all criteria at once, e.g. as in discrete choice experiments. Because AHP asks sequentially to consider two criteria out of a larger set, people could show a strong preference for criterion A if that criterion is compared with B, and a very weak preference for A if that criterion is compared to C. If subsequently B is compared to C, it could be that the respondents prefer B over C, even though this is not what would be expected based on the previous observations. However, because respondents did not have to rank all criteria at once, the scores on the AHP scales might come from a different underlying latent scale and might therefore not be directly comparable (e.g.

a +9 in a comparison between two criteria may only correspond to a +1 in the comparison between two other criteria).

For the analytic hierarchy process the act of responding is performed on a rating scale. Here, the same critiques are valid as for the rating exercises. On a large scale with 15 points, for instance, it is difficult to understand what exactly every point on the scale means when making a choice.

Studies that used this technique found that mainly in face-to-face interviews there was a high level of acceptability. Some authors further found that people with a limited educational background did not have difficulties understanding the questions.¹²

This method can also be employed with computer programs for data collection purposes. This makes it easier to calculate the weights directly and offers the possibility to show the weights along the progress of the respondents' answers. The computer program can also use consistency tests such as transitivity. High levels of consistency have been reported in general.¹²

Best-worst methods

There were no studies in our systematic review that used a best-worst scaling (BWS) method. This method has been identified through hand searching and expert advice. We found a number of papers that assessed the performance of best-worst scaling methods and compared them to other methods. The evidence is mixed. Some authors find important differences in attribute weights between DCE and best-worst scaling,^{83, 84} while others find the attribute weights to be quite similar after rescaling.^{85, 86} Whitty et al. (2013) compared best-worst scaling and discrete choice experiments using empirical data and found that both techniques generally generated consistent results in terms of the strength of preferences for certain attributes. For example, respondents showed stronger preferences for preventive technologies or early diagnosis than for technologies improving quality of life, reducing side-effects or reducing waiting times. However, the relative attribute weights differed between DCE and BWS for some attributes (e.g. benefit and age) and as a consequence also the preference orderings of scenarios consisting of a combination of levels of attributes. People rated the DCE task as less difficult than the best-worst scaling task. The majority of the respondents preferred the DCE over the best-worst scaling task.⁸³



This was also a finding in a study by Xie et al. (2014), investigating the possibility to elicit preferences for the EQ-5D-5L^b using BWS or DCE.⁸⁴ The authors also found that the intra-class correlation coefficient, used to test the test-retest reliability of both techniques, was higher for the DCE than for the BWS.

This contrasts with the results of a study by Potoglou and colleagues (2011), who found that the preference weights obtained by means of DCE and BWS do – in most cases – not differ significantly after rescaling.⁸⁵ Similar findings were described by Severin et al. (2013)⁸⁶. As Whitty et al. (2013), both Severin et al. (2013) and Potoglou et al. (2011) also found similar patterns in preferences.

Best-worst scaling seems to be easier than ranking. Best-worst methods are believed to give more incentives to respondents to discriminate between alternatives than ranking and rating exercises. They ask for the most important and the least important alternative and do not allow for any middle ground.⁷⁸

4.3.4 Discussion and conclusion

Comparing the validity of ranking, rating and choice-based methods on the basis of empirical research applying one of these techniques is difficult because studies use different criteria, different wordings, and different data collection interfaces, which all interact with the validity, reliability and internal consistency of the results. Some studies tried to test the validity of the method by using a second or third method in order to look at differences in outcome. Some found similar results, others found inconsistencies.

Acceptability is a subjective criterion and depends for instance on the educational level of the respondents. The acceptability could also depend on the amount of information given to the respondents. The information could be structured in several ways. Some researchers provide a comprehensive introduction giving information about the need to ration health care, others merely make respondents aware of the fact that there is

a limited budget for health care that needs to be spent according to general principles, still others provide ad-hoc information in every step of the survey, or even only if the respondents ask for it.

Overall there was very little information about reproducibility of any of the preference elicitation techniques.

Table 11 summarises some strengths and weaknesses of different techniques.

When deciding which method to use, it is important to consider the possible hurdles or insights a specific method can provide relative to the aim of the study. Statistical efficiency and response efficiency need to be weighed against each other. The higher the statistical efficiency and the more built-in checks for the reliability and validity of the data, the longer the questionnaire and the lower the response efficiency. Respondents will disconnect when the burden is too high.

Economists seem to favour DCE, because it is based on the utility theory, which is the foundation of classical welfare economics. DCE has the important merit of making opportunity costs visible and explicit to respondents, thereby representing more closely the decision problem policymakers are facing. Research with DCE is extensive, but translation towards real-life decision making is often still missing. We have chosen DCE for the current study based on the theoretical foundations of DCE and on the empirical evidence with regard to its acceptability. We considered that we could obtain a good balance between statistical efficiency and response efficiency by limiting the number of choice sets/questions to less than 9 per person (including a choice set with one dominant alternative to check for consistency) and distributing 24 different versions of the questionnaire (see paragraph 5.1.9 where we describe the design of our DCE). A further advantage was that DCE could be administered as a web-based questionnaire.

^b EuroQol health-related quality of life instrument with five dimensions and five levels per dimension



Table 11 – Overview of strengths, weaknesses and limitations of different preference elicitation techniques

	Ranking	Rating	Simple choice	DCE	AHP	BWS	Pie method
Strengths	Allows the respondent to rank all elements Simple data analysis	Relative simple exercise Strength of preference	Simple exercise Potential to be realistic	Complex choices, as in real life Info about trade-offs and opportunity costs	Both exercise about abstract criteria and concrete cases	Easier than grand ranking exercise, because smaller exercises	Relatively easy exercise Clear opportunity costs Good reproducibility
Weaknesses	Often more abstract exercises, less realistic If a long list of elements, could be difficult exercise	Difficult to know how scale is interpreted by respondent (intervals)	If choices per criterion, the respondent does not have a lot of information and loses overview	Much information is retrieved from responses but a clear scope is needed to interpret results Complex data analysis	Because answers only relate to one characteristic there is no overview for respondent of general ranking of criteria	Outcome is not much better than ranking outcome of abstract elements, no concrete information	Respondents may feel uncomfortable with the exercise if framed as allocation of a certain budget (monetary terms)
Limitations	No information about equal preferences or opportunity costs Difficult to use results for concrete policy making	No info about opportunity costs				No opportunity costs or trade-offs visible	No clear info about choices or trade-offs between attributes

DCE=discrete choice experiment; AHP=analytic hierarchy process; BWS=best-worst scaling



Key points

- Ranking, rating and choice-based techniques can be applied in stated preferences approaches for eliciting quantified preferences from citizens.
- No single technique stands out in all respects as the best approach for measuring preferences about reimbursement criteria. There is large variability in how the techniques are applied, which complicates the assessment of the validity, reliability, internal consistency and acceptability of the techniques. Most important is that the preference weights serve the final purpose of the exercise and withstand the reasonableness check.
- Discrete choice experiments (DCE) are based on random utility theory and therefore preferred by economists. We have chosen this technique for the current study because it requires respondents to consider several criteria at once, similar to what decision makers are supposed to do in real life. Moreover, as research demonstrated good acceptability of DCE, we presumed to be able to obtain a good balance between response efficiency and statistical efficiency with a design that would require respondents to answer 9 choice questions, including a question that included a dominant alternative.

5 SURVEY ON THE RELATIVE IMPORTANCE OF DIFFERENT DECISION CRITERIA FOR REIMBURSEMENT

5.1 Methods

For the measurement of the relative importance of different decision criteria for the reimbursement of health interventions according to the general public, we performed a large survey in a representative sample of the general population. The same survey was sent to a large group of stakeholders currently involved in decision-making processes on the federal governmental level in Belgium. We followed the practice guidelines for survey research published by the Ministry of the Flemish community to define the different phases of the project.⁸⁷

5.1.1 Choice of data collection technique

We chose to perform a web-survey, giving also the possibility to request a paper version of the questionnaire. Web-surveys have the advantage that they can contain built-in dependencies and required fields.⁸⁸ Questions on which a response is absolutely needed for the study purposes are typically conceived as required fields, i.e. if no answer is provided, the software does not move to the next question but gives a message that the question should be answered to be able to move to the next question. We applied this to all choice questions (see infra). Built-in dependencies are questions asked conditional upon a previous response. For example, when asked to make a choice between two scenarios, we asked people to indicate how certain they were of their choice. Only respondents who answered “very uncertain” or “uncertain”, were asked to state why they were uncertain. The “reason for uncertainty”-question is an example of a built-in dependency.

In the paper version of the questionnaire these techniques cannot be applied. Therefore, the risk of unusable questionnaires is higher. However, the type of questions did not allow for telephone survey.

Web-based questionnaires are more practical and budget-friendly. According to a survey from the Federal Public Service Economy, 84% of Belgian citizens used internet in 2013.³ It is therefore expected that the majority of the population would be able to participate in a web-survey,



although a bias towards specific population groups cannot entirely be excluded. The risk of responder bias is a general problem with every approach used, however, and can only be reduced by combining several survey techniques. We analysed the difference in characteristics of people who responded electronically and people who responded on paper. In addition, we examined whether people who responded immediately after the first invitation, after the first, second or third reminder had different preferences. The hypothesis is that preferences of late responders (those responding after a couple of reminders) are more likely to resemble those of non-responders than those of people responding immediately after the initial invitation.⁸⁸ Of course, this is a hypothesis, for which there is – for obvious reasons – no evidence.

Face-to-face interviews would have been very costly if we wanted to reach a representative sample of the general public in Belgium. Moreover, it is uncertain whether the quality of the data would have been much better with face-to-face interviews than with a web-based survey. The objective of the survey was not to explain the preferences. The exploration of possible reasons for making particular choices is the objective of the citizen-labs organised by the King Baudouin Foundation, subsequent to this survey. The citizen-labs are a qualitative approach that can help to understand the preferences of the public, the reasoning of respondents when considering several treatments and their weighting of pros and cons of different treatments. The citizen-labs will consist of three week-ends where citizens can interact with each other and with experts.

5.1.2 Questionnaire development process

A French and Dutch version of the questionnaire were developed by a native French-speaking researcher and a native Dutch-speaking researcher working together side by side and in continuous dialogue about the exact meaning of every single word used in the survey. The complete research team convened weekly to discuss any remaining open issues. Several rounds of revision were needed before a first version was ready for **pre-testing** in both languages. The objective of the pre-test was to verify the comprehension of the questions in a selection of people. Every researcher of the team selected a number of people, with different socio-educational and socio-economic backgrounds, to fill out the questionnaire and discuss whether the questions were clear, whether they thought the survey was

feasible and how they made their choices. A checklist of questions to ask and a process note was created to guide the researchers through the pre-test interview (see Appendix). During the interview, the respondent was asked to read each question out loud. The researcher systematically asked for feelings, comprehension of the questions, how the respondent thought he or she was supposed to deal with the question and response options, and whether the explanations provided in the pop-ups were sufficiently clear and helpful. The pop-ups are explanations of difficult words used in the questionnaire, which are provided as footnotes in the paper version and as pop-up boxes in the web version when the respondents puts his or her cursor on the word. Respondents were also invited to explain how they made their choice and which reflections they made when making their choice in the choice questions.

Twenty-two pre-test interviews were made. The team met two times to present and discuss the results of all pre-test exercises. The team subsequently agreed on solutions to problems of clarity or adaptations to improve acceptability and consistency. Modifications related to the framing of some questions, clarification of concepts and in one question the reduction of the number of response options to rank (moral reasoning exercise).

The modified versions of the questionnaire were transferred into a web-environment. We used LimeSurvey, an open source survey application, to develop the web-versions. Invitations to participate in the pilot test were accompanied by an information sheet about the questionnaire, similar to the information sheet for the general public who would receive the invitation.

A **pilot test** of the survey was performed in 219 people, selected amongst the relatives and friends of employees of KCE. The objective of the pilot survey was to obtain more feedback on the feasibility, readability and comprehension of the questionnaire. Employees were asked to select respondents from all educational levels whenever possible. People who participated in the pilot survey were asked, at the end of the survey, to comment on (1) the formulation of the questions (clarity, comprehension), (2) the definitions of difficult words or words with a specific meaning and (3) the lay-out and form of the survey (font size, colours of text and background, readability). There was also space for free comments. The objective of the pilot survey was *not* to derive relative preferences for different decision criteria. This was not feasible, because the design of the survey requires 24



different versions of the questionnaire to be completed by a sufficiently large number of people (at least 1000) to be able to derive meaningful results from the survey results (see section on “survey design”). Expecting still a significant number of comments on the pilot version of the questionnaire, requiring additional modifications before starting the survey in the general public, and the absence of representativeness of the population filling out the pilot questionnaire, the relative weights of criteria obtained from the pilot survey would not have been meaningful or useful for comparisons anyway.

A **test-retest** reliability check was performed in 42 people. Employees of KCE and students of the UHasselt participated in the test-retest surveys. The retest was performed on average two weeks after the test. The objective of the test-retest reliability check was to assess the stability of the answers and hence the reliability of the questionnaire.

5.1.3 Sample selection

A representative sample of 20 000 people from the Belgian general population, stratified by age and sex, was drawn by the National Registry. All people with a national number in the Belgian National Registry between 20 and 89 years of age were eligible for participation. The invited sample thus consisted of about 1 out of every 430 persons of the Belgian population within the eligible age group. The selection started at the age of 20 to be able to select people up to the age of 89 and at the same time obtain seven categories with an equal age range (20-29, 30-39, 40-49, etc.). If the response rate across age and sex differs significantly from the expected response rate per category based on the stratified sample, we'll check the need to weigh by age and gender in the analysis.

Stratification by age and sex was considered important because literature suggests there is a relationship between preferences and age and sex. Other criteria that might determine preferences (e.g. being ill, being religious, etc.) cannot be included as sample selection criterion because this information is not available in the database of the National Registry. Place of residency is not included as a selection criterion, implying that place of residency is not assumed to be correlated with preferences regarding priority setting criteria. This may be a too strong assumption because place of residency may be correlated with level of income or education, which may in turn be correlated with preferences. However, the question then becomes which level of detail is relevant and should be envisaged if place of residency

is to be included as a selection criterion. Every choice, from large to very small geographical entities, has weaknesses and strengths, but the effort and cost of including place of residence as a selection criterion might not be worth the benefit in terms of the representativeness of preferences we aim for. Uncertainty will remain. For the selection of a sample, an approval by the Privacy Commission (Sectoral Committee of the National Registry) had to be obtained. This has been provided on 3 February 2014, based on the deliberation of 11 December 2013 (RR 77/2013).

In addition to the sample from the general public, we invited all members of nine decision-making or consultative bodies in health care on the national level, to fill out the survey. Their answers were analysed separately from those of the general public. A total of 421 members of decision-making or advisory organs were invited to participate in the survey. They included decision-making organs of the National Institute for Health and Disability Insurance, the Federal Public Service Public Health, the policy unit of the minister of Public Health, the Chamber of Representatives and the Senate and an independent advisory committee.

The following advisory commissions and decision-making bodies received an invitation by email:

- The Belgian Advisory Committee on Bioethics (Belgisch Raadgevend Comité voor Bio-ethiek / Comité consultative de Bioéthique de Belgique, independent advisory committee)
- The Drug Reimbursement Committee (Commissie Tegemoetkoming geneesmiddelen / Commission de Remboursement des Médicaments, RIZIV / INAMI)
- The Committee for the Reimbursement of Implants and Invasive Medical Devices (Commissie Tegemoetkoming Implantaten en Invasieve Medische Hulpmiddelen / Commission de Remboursement des Implants et des Dispositifs Médicaux Invasifs, RIZIV / INAMI)
- The college of medical doctors-directors (College van geneesheren-directeurs / Collège des médecins-directeurs, RIZIV / INAMI)
- The Technical Medical Council (Technisch Geneeskundige Raad / Conseil Technique Médical, RIZIV / INAMI)



- Federal Agency for Medicines and Health Products (Federaal Agentschap voor Geneesmiddelen en Gezondheidsproducten / Agence fédérale des médicaments et des produits de santé)
- Senate's commission of social affairs (Commissie sociale zaken / Commission des affaires sociales)
- Chamber's commission public health (Kamercommissie Gezondheid / Commission Santé publique)
- Policy Unit of the minister of Public Health (Beleidscel Minister Volksgezondheid / Cellule stratégique Ministre de santé publique)

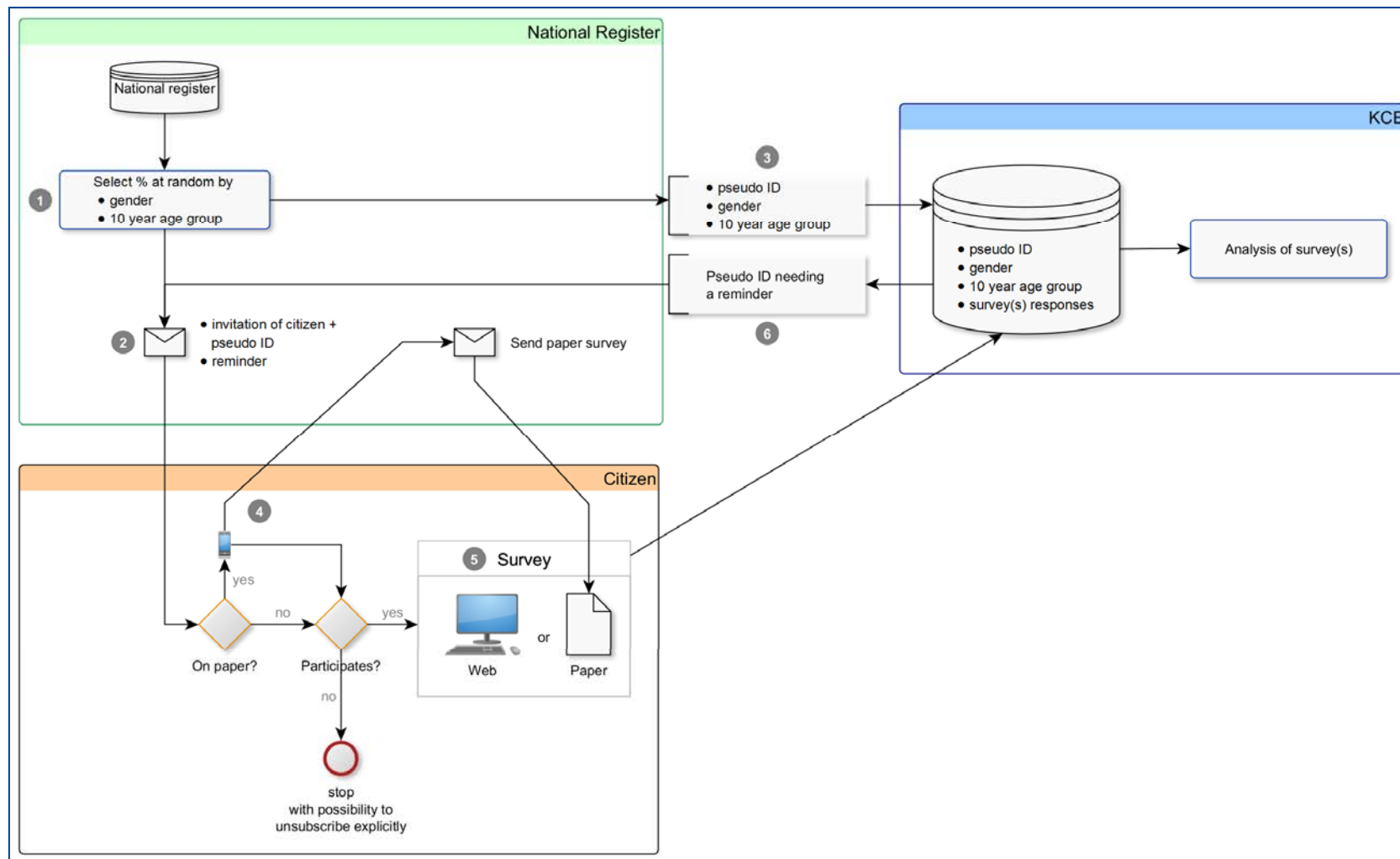
5.1.4 *Anonymity*

The survey was performed anonymously. To ensure anonymity, KCE first developed a list of unique codes and sent this list to the National Registry. The National Registry allocated a code to each individual drawn randomly from the population registry and sent the invitation letters by regular mail on behalf of KCE. Responses arrived directly at KCE with only the respondent's code and not his/her address. The data and information streams are presented in Figure 13.

People who preferred to fill out a paper questionnaire were asked to call the National Registry. The information on how to request a paper version was provided in the invitation letter. When a person called, the National Registry then asked for the unique code of the respondent, wrote this on the questionnaire and sent a paper version of the questionnaire to the respondent, together with a pre-stamped envelope addressed to KCE. As such, also the paper responses arrived directly at KCE, with only the respondent's code and not his/her address.



Figure 13 – Survey process





5.1.5 *Invitation letters and reminders*

Twenty thousand people received a letter from KCE, inviting them to participate in the survey. The letter explained the objective of the study and referred to a web-page where people could log in using their personal code mentioned in the letter.

Three reminders were sent to non-responders, each one two weeks after the previous mailing as recommended in several methodological reference documents.⁸⁸ KCE provided the codes of respondents to the National Registry to let them know who should not receive a reminder. The initial invitation letter as well as the reminders contained more detailed information about the survey.

The letters contained information on:

- the objective of the study;
- the importance of participating;
- the respondent's unique personal access code;
- the anonymity of the survey;
- how to request a paper version of the questionnaire;
- informed consent;
- the citizen's laboratory organised by the King Baudouin Foundation.

It was also explained that by participating, people consented to analysis of their responses but that no individually identifiable responses would be made public. The invitation and reminder letter and the complementary information sheets included in the mailings can be found in appendix.

5.1.6 *Preference elicitation technique*

Based on the overview of the advantages and disadvantages of different preference elicitation techniques (see 4.3), the discrete choice experiment (DCE) technique was chosen for collecting the data on relative preferences in our survey. DCE is considered to be a difficult technique but possibly fitting best our purpose of deriving relative preferences for different criteria. This technique is well embedded in economic theory and resembles best the actual difficulty of making decisions. Compared to for example a rating and ranking exercise, DCE has the advantage to let all attributes be weighted in

the decision at once. In theory, this should maximize the information that can be obtained from the respondent's choices.

In DCE, respondents are asked to choose between two or more different hypothetical alternatives, where each alternative is described by a set of attributes (characteristics). The attributes are the same for the alternatives, only the levels differ between them. Based on the choices people make when confronted with different choice sets, one can determine the relative importance of each attribute in the choice. For example, current life expectancy and quality of life could be two attributes of therapeutic need. If you want to know the relative importance of life expectancy as compared to quality of life for defining therapeutic need, you could present a number of choice sets to respondents where the levels of life expectancy and quality of life differ between the scenarios among which the respondent has to choose. Based on the choices he/she makes, it will become clear which attribute is most important for the decision.

Because we were solely interested in the relative importance of the attributes, we used an unlabelled DCE.⁸⁹ In unlabelled DCE, the labels of the alternatives have no meaning in themselves (e.g. alternative 1 and 2). This implies that trade-offs are fully dependent on the attribute levels.

We used a DCE with two alternatives to choose from. An issue with DCE is that the number of possible choice sets increases exponentially with the number of attributes and levels. Obviously, the higher the number of attributes and the higher the number of levels per attribute, the higher the number of possible scenarios. With four attributes and (only) three levels within each attribute, 81 different scenarios can be described and mathematically 6561 binary choice sets are possible, including sets that would never be included in an actual DCE such as scenarios that are compared with themselves, or sets that are the same in content but different in order of presentation (left and right). There are $80 \times 81 = 3240$ binary choice sets that differ in content. It was obviously impossible to present all possible choice sets to each respondent. Therefore, the number of choice sets per respondent had to be reduced. This was done according to experimental design principles (as described in 5.1.9).



5.1.7 *Criteria*

As explained in section 1.1, the focus of this study is on reimbursement criteria, rather than on general ethical principles for reimbursement. Given the typical incremental nature of reimbursement decisions in health care, it is important to also put the criteria into this context. The alternative to which a new intervention is compared is not necessarily an active treatment. It can be “no treatment” or “best supportive care” if no alternative treatment exists.

The criteria included in the survey were therefore always defined “given the currently available treatment or care” or “compared to the currently available treatment or care”, depending on whether the question related to the need for a better treatment, or the added value of a new treatment respectively.

The questions of our survey were structured in three blocks, corresponding to question 1 and 3 of the KCE MCDA framework (see Table 6: therapeutic need, societal need and added value of treatment. We did not include medical need, which refers to the severity of illness as reflected by its impact on patients if no treatment would be provided. As our focus is on incremental decisions, therapeutic need is more relevant than medical need. A high medical need may correspond to a low therapeutic need if an effective treatment is already available for the disease.

The attributes and levels were operationalised as presented in Figure 14. These operational descriptions were used to construct the DCE scenarios in each block.

**Figure 14 – Blocks, attributes and levels used in the survey****Therapeutic need: characteristics of the health condition, given current standard treatment**

- Discomfort / inconvenience of current treatment
 - Patients ...
 - experience much discomfort from current treatment
 - experience little discomfort from current treatment
- Quality of life with current treatment
 - Patients ...
 - currently have a quality of life of 8 out of 10
 - currently have a quality of life of 5 out of 10
 - currently have a quality of life of 2 out of 10
- Life expectancy
 - Patients ...
 - no longer die from the disease
 - die 5 years earlier than people without the disease
 - die almost immediately from the disease, despite current care
- Age group
 - Patients ...
 - are older than 80 years of age
 - are between 65 and 80 years of age
 - are between 18 and 64 years of age
 - are younger than 18 years of age

Societal need: characteristics of the illness

- Prevalence
 - The disease ...
 - is rare: less than 2000 people in Belgium have the disease
 - is not so frequent: between 2000 and 10 000 people in Belgium have the disease
 - is rather frequent: between 10 000 and 100 000 people in Belgium have the disease
 - is very frequent: more than 100 000 people in Belgium have the disease
- Disease-related public expenditures
 - Little public expenditures per patient
 - Much public expenditures per patient



Added value of the new treatment compared to the existing alternative

- The new treatment, compared to the already available treatment ...

- *[impact on the discomfort of treatment]*
 - reduces the discomfort of treatment for the patient
 - gives as much discomfort to the patient
 - increases the discomfort of treatment for the patient
- *[impact on quality of life]*
 - improves the quality of life of patients
 - does not change the quality of life of patients
 - reduces the quality of life of patients
- *[impact on life expectancy]*
 - does not change the life expectancy of patients
 - increases the life expectancy of patients
- *[impact on the prevalence of the disease]*
 - cures fewer patients
 - cures an equal number of patients
 - cures more patients
- *[impact on public expenditures]*
 - reduces the disease-related public expenditures per patient
 - does not change the disease-related public expenditures per patient
 - increases the disease-related public expenditures per patient

Discomfort of treatment was included because it is one of the decision criteria defined by Royal Decree in Belgium.⁹⁰ It refers to the inconvenience of a treatment, caused by for instance frequency of use (e.g. taking a drug once or more times a day), the administration route (e.g. syringes, oral drugs, administration by yourself or by someone else), the place of administration (at home, in the hospital, in a doctor's cabinet).

Life expectancy and **quality of life** are both components of the QALY concept. Quality of life and life expectancy were included as separate criteria to allow people to make a trade-off between living long and living good. Quality of life was described as having five dimensions (as in the EQ-5D): mobility, self-care, usual activities, physical pain and mental suffering (anxiety/depression).

Age, although not considered to be a criterion as such, was added to put "impact of disease on life expectancy" into perspective. It is difficult for respondents to judge the importance of losing 5 years of life expectancy without knowing the age of the patient. We wanted to leave open the option of having different preferences for life expectancy depending on age and therefore included age as a descriptive fact. This had no impact on the design of the questionnaire, but had to be taken into account in the analysis of the data.

The **frequency of disease** was included because we wanted to know to what extent the "rarity" as such is an important criterion for defining needs, independent from the severity of the disease. It refers to prevalence of the disease.

The **disease-related public expenditures** were defined as the total public expenditures per patient with the disease, including health care expenditures and productivity losses leading to benefits that replace wage losses, etc.



The **attributes for added value** correspond to the improvement on the attributes of therapeutic and societal need. The rationale behind this choice was that if criteria were considered relevant for determining therapeutic and societal need, improvements on these criteria would also be relevant for determining the added value of a new treatment. In contrast to the needs-assessment, where the individual patient-related criteria (in therapeutic need) were considered separately from the society-related criteria (in societal need) to keep the task manageable and clear, all attributes were considered at the same time in the added value assessment. The attributes in the added value assessment were all related to the impact of the new treatment, hence treatment-related attributes. For the clarity and feasibility of the questionnaire, it was considered useful to make this distinction between condition-related individual attributes, condition-related societal attributes and treatment-related attributes.

The attributes for therapeutic and societal need both described the situation of a patient group or illness with current treatment. There is no comparator involved in these attributes. This contrasts with the attributes for added value, which describe how a new treatment improves the attributes compared to current treatment.

5.1.8 Discrete choice experiment choice sets

Each choice set consisted of two alternatives between which the respondent had to choose. The alternatives did not represent a specific disease or treatment but were generic.

- In the first block respondents were asked to choose the scenario in which the need for a better treatment than the one that already exists is the highest from the patient's point of view (see Figure 15 for an example of one DCE question on therapeutic need). The scenarios related to different patient groups.
- In the second block respondents were asked to choose the scenario in which the need for a better treatment than the one that already exists is the highest from the societal point of view (see Figure 16 for an example of one DCE question on societal need). The scenarios related to different diseases.
- In the third block respondents were asked to choose the new treatment that he/she would most prefer to be reimbursed, if the two treatments were for the same disease. The one that the respondent chose is considered to have the highest added value according to the respondent (see Figure 17 for an example of one DCE question on added value of treatment).

**Figure 15 – Example of a DCE question for therapeutic need**

Example: Two patient groups are described below. Both patient groups currently receive treatment. The discomfort associated with the treatment, the quality of life and life expectancy of patients getting this treatment and the typical age of patients with this condition are as follows:

Patients of group 1	Patients of group 2
<ul style="list-style-type: none">• have a quality of life of 8 on 10• experience much discomfort from treatment• are between 18 and 64 years of age• no longer die from the disease	<ul style="list-style-type: none">• have a quality of life of 5 on 10• experience little discomfort from treatment• are older than 80 years of age• no longer die from the disease

For which patients do you consider it most important to develop a new and better treatment? You may define yourself what you consider to be “better”.

Choose one group.

<input type="checkbox"/> Patients of group 1	<input type="checkbox"/> Patients of group 2
--	--

Figure 16 – Example of a DCE question for societal need

Example: Two diseases are described as follows:

Disease 1	Disease 2
is not so frequent: between 2000 and 10 000 people in Belgium have the disease. Every patient costs little to society.	The disease is rather frequent: between 10 000 and 100 000 people in Belgium have the disease. Every patient costs much to society.

For which disease do you consider it most important to develop a new and better treatment? You may define yourself what you consider to be “better”.

Choose one disease.

<input type="checkbox"/> Disease 1	<input type="checkbox"/> Disease 2
------------------------------------	------------------------------------

**Figure 17 – Example of a DCE question for added value of treatment**

Example: Suppose two new treatments appear on the market for the same disease. There is already a treatment available for this disease, which is fully reimbursed by the health insurance.

You can decide yourself which of the two new treatments the health Insurance should reimburse.

There is only enough money to reimburse one of the two treatments. Patients who wish to receive the treatment you do not choose, will have to pay for this themselves.

New treatment 1	New treatment 2
<p>The new treatment, compared to the already available treatment,</p> <ul style="list-style-type: none">• gives as much discomfort to the patient• does not change the quality of life of patients• reduces the cost of each patient to society• cures an equal number of patients• increases the life expectancy of patients	<p>The new treatment, compared to the already available treatment,</p> <ul style="list-style-type: none">• increases the discomfort of treatment for the patient• improves the quality of life of patients• increases the cost of each patient to society• cures fewer patients• does not change the life expectancy of patients
<p>In your opinion, which treatment should be reimbursed? Choose one treatment.</p>	
<input type="checkbox"/> New treatment 1	<input type="checkbox"/> New treatment 2



5.1.9 Design of DCE choice sets

If no practical limits are taken into account, a full factorial design using each combination of the levels of all attributes would be ideal. It would allow the estimation of all main attribute effects and all interaction effects between attributes. However, this becomes unfeasible when the number of attributes and levels rises.⁹¹ Even a fractional factorial design, i.e. a reduced design where higher-order interactions (e.g. three-way and up) can no longer be estimated unconfounded, might prove too large in terms of different survey versions and required sample size to be feasible. Following Reed Johnson et al (2013), we took the following general design issues into account:

- Dominant choice sets: in the Therapeutic need and Added value domain, the levels are assumed to have a certain order. A dominant choice set then is a set where the levels of the attributes of one alternative are systematically higher ranked than those of the other alternative. These were removed from the set of possible choice sets (design space).
- Overlap: when for the same attribute the levels are the same between the two alternatives, there is overlap between the choice sets. Because there are only two attributes in the Societal need domain, no overlapping choice sets were retained. For the Therapeutic need and Added value domain, a maximum overlap of one attribute was allowed for.
- Implausible combinations: in each domain, there was a check on combinations of attribute levels that are implausible or illogical. None were found.
- Balance between response efficiency and statistical efficiency: statistical efficiency is about minimizing the confidence intervals around the parameter estimates of the model given a particular sample size. Response efficiency is about measurement error due to factors related to the respondent, like fatigue, inattentiveness, non-comprehension. The two are often in conflict about number of choice sets per respondent or number of attribute combinations. Given the more abstract, general nature of the topic, we favoured response efficiency over statistical efficiency.

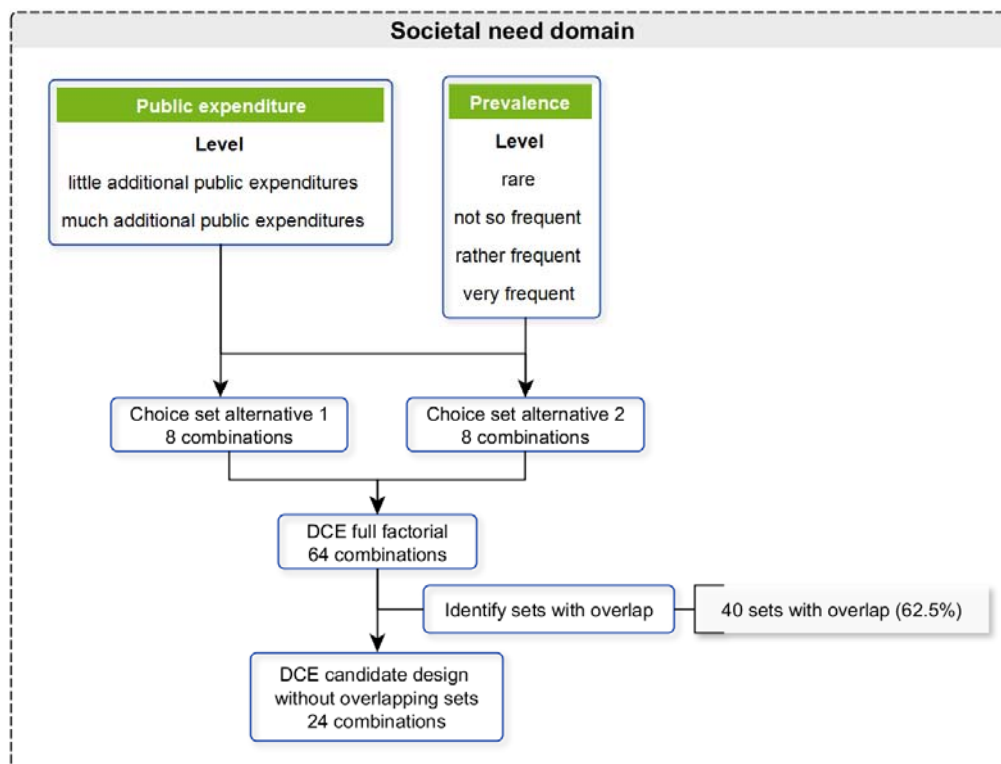
The following sections describe in detail the design choices made per domain.

5.1.9.1 Societal need domain

In the Societal need domain, the respondent was asked to choose between two diseases for which it is the most important to develop a new and better treatment. The diseases were defined in the choice sets by two attributes: the per-patient public expenditure related to the disease and the prevalence of the disease. The attribute public expenditure had two levels: little and much. The attribute prevalence had four levels: less than 2000 patients, between 2000 and 10 000 patients, between 10 000 and 100 000 patients, and more than 100 000 patients. Combined in a choice set, this gives eight different sets. The DCE with two choice sets then has 64 different combinations, of which 40 combinations have at least one overlapping attribute between choice sets. This rule also excludes the choice sets that have the same levels per attribute in each set (i.e. all attributes overlap). The remaining 24 combinations consist of 12 different choice sets, each used twice (set A for alternative 1 and set B for alternative 2; set B for alternative 1 and set A for alternative 2) (see Figure 18).



Figure 18 – Choice set design in the Societal need domain





5.1.9.2 Therapeutic need domain

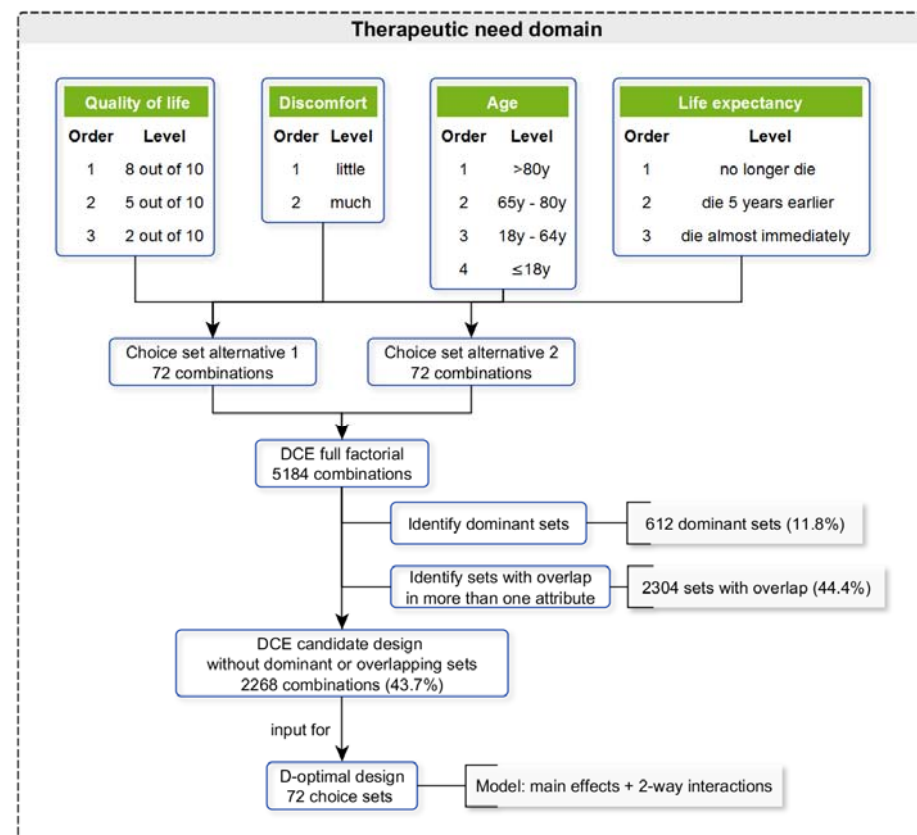
In the Therapeutic need domain, the respondent was asked to choose between two groups of patients for which it is most important to develop a new and better treatment. The groups of patients were defined by four attributes: quality of life, discomfort, age, and life expectancy (for the levels, see Figure 19). After discarding dominant choice sets and choice sets with more than one attribute overlapping, we retained 2268 combinations as design space. From these, 72 combinations were determined by searching for a D-optimal design for a main effects and 2-way interactions model (see Figure 19). A D-optimal design allows to estimate the parameters of the model without bias and as precise as possible, given the constraints on the number of choice sets and the chosen statistical criterion. We chose the D-criterion, the determinant of the information matrix ($|X'X|$ with X being the model matrix of the parameter estimates). The algorithm for D-optimality maximizes the D-criterion.⁹² In our case, we estimated non-linear models (see below in 5.1.13).

For non-linear models, the efficiency or optimality of the design does not only depend on the information matrix, but also on the unknown parameter estimates.⁹³ Because we had no prior information on the relative preferences of our chosen attributes we assumed zero prior parameter values.

As an alternative, we could have applied a Bayesian approach to the D-optimal design⁹⁴. This means we would use available information on likely parameter estimates to specify a proper prior distribution of the parameter estimates as input to the design process⁹¹. If the informative prior distribution is correctly specified, the Bayesian algorithm to construct choice sets would avoid dominant choice sets and would result in more precise parameter estimates than with the current assumption of zero prior parameter values, given an equal number of respondents. However, we felt that the prior information available on the preferences for the chosen attributes levels was insufficiently univocal and backed by scientific evidence to construct an informative prior distribution. Taking this into account and the fact that the difference between our approach and a Bayesian approach would mainly influence precision of our estimates and not possible biases, we opted to use the design approach described in the previous sections.

The resulting 72 combinations gave 24 versions with three choice sets randomly assigned per version.

Figure 19 – Choice set design Therapeutic need domain



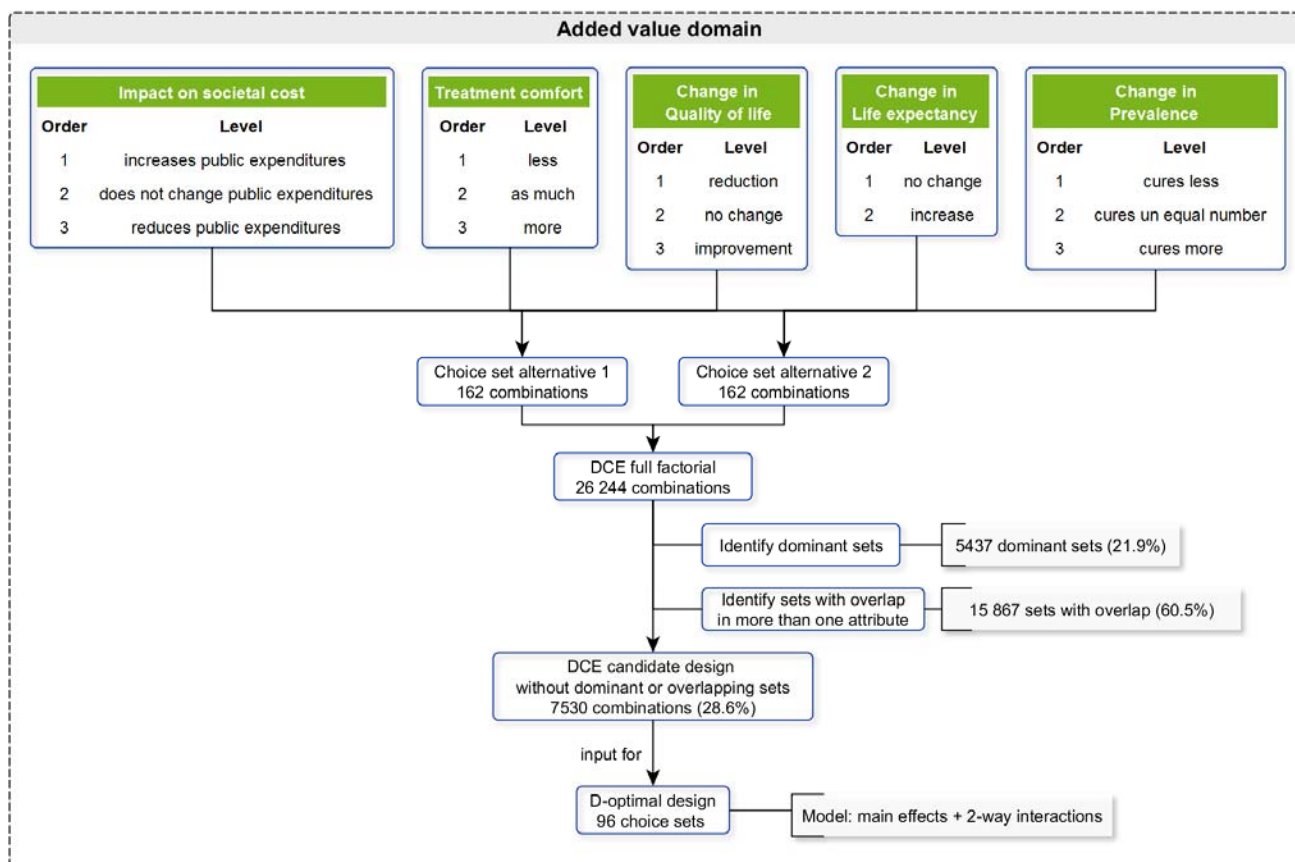


5.1.9.3 *Added value domain*

In the Added value domain, the respondent was asked to choose the new treatment out of two for the same disease that deserves most to be reimbursed by the health insurance. The treatments were defined by five attributes: change in quality of life, change in life expectancy, treatment comfort, change in prevalence, and budget impact (see Figure 20). After discarding dominant choice sets and choice sets with more than one attribute overlapping, we retained 7530 combinations as design space. From these, 96 combinations were determined by searching for a D-optimal design for a main effects and 2-way interactions model (see Figure 20). These 96 combinations gave 24 versions with 4 choice sets randomly assigned per version.



Figure 20 – Choice set design Added value domain

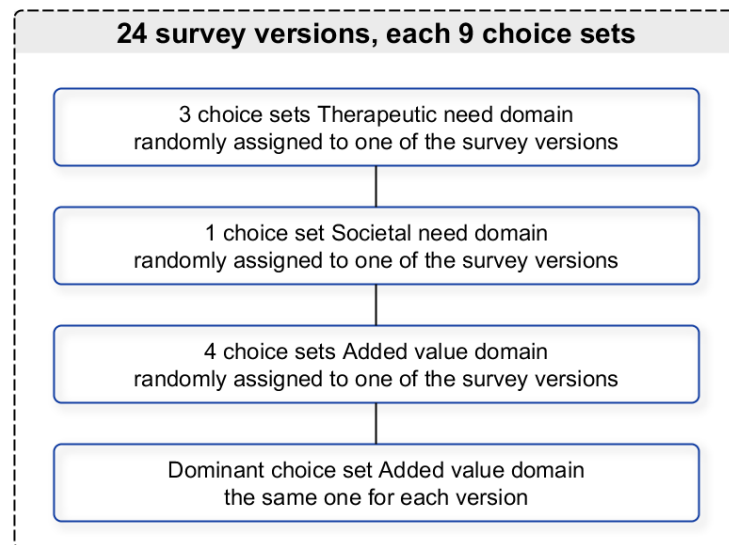




5.1.10 Survey versions

Combining the Therapeutic need domain, Societal need domain, and Added value domain choice sets resulted in 24 different versions of the survey. Additionally, one dominant choice set was added to be able to check comprehension and attentiveness in respondents. Each survey version consisted of three Therapeutic need domain choice sets, one Societal need domain choice set, four Added value domain choice sets and a dominant choice set from the Added value domain (see Figure 21). The dominant choice set was put in the middle of the Added value domain choice set block.

Figure 21 – Survey versions



5.1.11 Survey completion and response registration process

The twenty-four different versions of the questionnaire were encoded in LimeSurvey. When respondents entered the link to the web-survey mentioned in their invitation or reminder letter, they arrived at a welcome page, where they had to enter their personal code and choice of language to be able to continue. The welcome page also allowed for signing out of the survey. As soon as respondents had filled out their personal code, they were encoded as responders.

To have as representative responses as possible on each of the 24 versions, people of the same sex and age category received versions in the order of logging into the system. For example, the first man between 40 and 49 years of age received version 1 of the questionnaire, the second man between 40 and 49 years received version 2 and so on. With the 25th man between 40 and 49 years of age logging in the cycle restarted from version 1.

All choice questions were made obligatory, meaning that the respondent could not continue the survey if he/she did not respond to the question. It was possible to discontinue the survey and continue later on or to quit the survey altogether before completion. Incomplete responses were also encoded in the response database. Hence, in the description of the analysis sample, incomplete responses include both incomplete electronic responses and incomplete paper responses.

The members of the decision-making or consultative bodies received exactly the same questionnaire(s) as the general public. As for the general public, they entered the web-system and received one of the 24 versions according to the order in which they entered the system.

People requesting a paper version received either version 19, 20, 21, 22, 23 or 24.

Respondents who filled out the survey, either on the web or on paper, were asked at the end of the survey whether they would be prepared to be contacted for participation in the citizens' labs organised by the King Baudouin Foundation. The objective of the citizens' labs is to explore more in depth, with a small group of people who participated in the survey, the results of the survey, the current reimbursement decision-making criteria and possible improvements to these criteria. The participants to the citizens' labs could consult and hear experts.



5.1.12 Survey structure

The survey contained a set of demographic questions (such as age, sex, employment status, living situation, children Y/N, educational level), personal questions (general health level, severe disease Y/N, severe disease loved ones Y/N, affordability health care services), the choice sets for therapeutic need, societal need and added value and a moral reasoning exercise. The full questionnaire, in Dutch and French, can be found in appendix.

The demographic questions were split up. Part of the questions appeared at the very beginning of the survey, to have an easy start. Another part of the demographic questions was put at the end.

The moral reasoning exercise was basically meant as a “warming-up” exercise, to let respondents think about making choices in health care. It has been shown in literature that such a moral choice exercise improves the validity of the results of a discrete choice experiment.²⁶ Respondents were asked which kind of treatment they would give priority when having to decide about spending the resources of the health insurance to medical treatments. Six principles were presented. The list of principles was not exhaustive, of course, but because of the nature of the exercise, this was considered less important for our purposes. After making their choice of principles on which they would base their decisions, respondents were asked to order the principles they chose according to importance to them.

Each choice set was followed by a question about the level of certainty of the response. Respondents could indicate whether they were not certain at all, not certain, certain or very certain about their choice. This question was added for each choice question because the pilot survey showed that some respondents found it very difficult to make a choice. Not giving the opportunity to highlight that the choice made was not a very clear choice would likely increase the drop-out of the survey. For the choice sets included in the “added value”-block of questions, people who indicated being uncertain or very uncertain were moreover asked why they were uncertain. Three options were given, plus an open option “other”: (1) both treatments are equally good, it does not matter for me which one will finally be reimbursed, (2) none of the options deserves reimbursement and (3) the choice is difficult. Adding this possibility of specifying why the choice that had to be made (compulsory question) was uncertain, was again a result of

the comments received during the pilot test. Another option would have been to add a third response option “don’t know” next to the two alternatives between which the respondent had to choose, but this was considered “dangerous” for the design. It might have increased the number of undecided respondents significantly and therefore reduce the information on the basis of which an analysis could be made. Even people who feel only a little uncertain might be tempted to choose the “don’t know” option to opt-out, despite a (weak) preference for one of the two alternatives. The underlying assumption is now that people who are undecided have chosen randomly one of the two alternatives to be able to continue the survey.

5.1.13 Data analysis

All analyses were conducted in R 3.1.1, using the packages AlgDesign 1.1-7.2, car 2.0-21, lattice 0.20-29, mlogit 0.2-4, plyr 1.8.1, reshape2 1.4, sqldf 0.4-7.1 and vcd 1.3-1, in addition to the default packages.

Test-retest

The test-retest reliability was analysed with descriptive statistics and with Cohen’s Kappa as a measure of agreement.

Sample characteristics

Sample characteristics were analysed with descriptive statistics. The reported age category and gender was used in the analysis rather than the data available from the National Registry to take into account the possibility that the survey was completed by someone other than the invited. This seems to have been the case for 37 (0.8%) respondents with a different reported age category and for 159 (3.7%) with a different reported gender compared to National Registry data.

The comparison between the age and gender distributions of the sample and those of the general population was made with χ^2 tests.

Descriptive statistics were produced for several sub-groups of respondents, i.e. the group of respondents participating after initial invitation, after first reminder, after second reminder and after third reminder.



Relative preference weights analysis

Several methods to calculate attribute weights from the results of a discrete choice experiment have been described in literature.⁹⁵ Some of these methods require at least one quantitative attribute to ensure useful interpretation (e.g. price, survival in years or income for the calculation of marginal rates of substitution or willingness to pay). We used two different algorithms for the calculation of the relative weights per attribute: an algorithm based on differences in log-likelihood and an algorithm based on the range of the coefficients per attribute.

The following steps were taken:

1. Estimation of a multinomial logit regression – also referred to in literature as conditional logit – model:
 - a. The model contained only alternative-specific variables, representing the attributes of the choice sets. All main effects of the attributes were included in the model. The model has the general form of

$$P_{ij} = \frac{e^{(\beta \mathbf{X}_{ij}^T)}}{\sum_k e^{(\beta \mathbf{X}_{ik}^T)}}$$

where the dependent variable of the multinomial logit model represents the probability P that respondent i chooses alternative j out of k alternatives (two in this study). β is the matrix of estimated coefficients. \mathbf{X}_{ij}^T represents the transposed matrix of attribute values as presented to respondent i in alternative j (see point c below on the coding used for the attribute values). The coefficients for the model parameters were estimated by full information maximum likelihood method using the Newton-Raphson numerical optimisation routine⁹⁶. For the general population, each model was estimated a second time with a weight correcting for age and

gender distribution to correspond with the Belgian population. If the results are very similar, the unweighted models are used.

- b. No intercept was included in the model because the alternatives in our DCE were unlabelled.⁹⁷ For example, in Societal need, the diseases between which respondents had to choose were both only defined by their attributes, no other information was provided. Including an intercept would mean that the same attribute levels could have a different impact on the probability of choosing a disease. However, this would not make sense because the labels of the alternatives presented – “disease 1” and “disease 2” – have no meaning in themselves. Thus, trade-offs can only depend on the attribute levels, and not on other information meaning that the respondent has no other reason than the attribute levels to choose one or another disease. This is different in labelled DCE, where the label of the alternative can be part of the trade-off because it can represent additional information (e.g. instead of disease 1 and 2, lung cancer and pulmonary disease).^c
 - c. Effect coded contrasts rather than dummy coded contrasts were used for the model parameters of the attribute values.^{97, 98}. One advantage was that coefficient estimates and standard errors could be calculated for all levels of an attribute, because in effect coding all coefficient estimates for an attribute sum to zero. For the estimation process however, effect coding uses $n - 1$ levels per attribute (with n the number of levels of an attribute). We estimated the coefficient and standard deviation of the omitted attribute level but did not calculate the t-value as this is typically not explicitly part of the estimation process in case of effect coding.
 - d. The model fit was assessed by
 - i. comparing the observed proportions of alternative 1 and alternative 2 with the model predicted proportions of alternative 1 and 2.

^c It was our deliberate choice to exclude such implicit “additional information” from our design, because the additional information may be based on emotions, personal experiences, etc. which influence the choice

independently of the definition of the attributes. In that case, it becomes impossible to estimate the pure impact of the specific attributes. Because we want to develop a generic system, we need to have weights for generic attributes, independent of specific cases.



- ii. calculating the percentage of the choices correctly predicted by the model by comparing per choice set included in the 24 versions of the questionnaire the actual alternative chosen and the alternative with the largest probability of being chosen as predicted by the model.
- e. As the model coefficients estimated as such are not easily interpretable, we used two derived units:
 - i. Utility values: for each scenario (i.e. a particular combination of levels for each attribute), the coefficients of the levels were summed. Higher values reflect a higher therapeutic need, higher societal need or higher added value, depending on the domain.
 - ii. Probability of choosing a specific scenario out of the full set of scenarios that can be described by combining levels of different attributes:

$$P_{set} = \frac{e(\sum C_{ij})}{\sum_k e(\sum C_{ij})}$$

with P_{set} the percentage of respondents that would choose a particular scenario given the choice between k other scenarios. $\sum C_i$ is the sum of the coefficients per the chosen level of attribute.

2. Calculation of relative preference weights using the log-likelihood method:
 - a. Calculate the log-likelihood for the model.
 - b. Calculate the log-likelihood for the model *minus* one of the alternative specific variables, which represents the attribute of interest (=the reduced model).
 - c. Test if the reduced model is statistically equal to the full model with the likelihood ratio test. If the test rejects the equality hypothesis, consider the relative importance of the removed attribute to be different from zero.
 - d. Calculate the difference in log-likelihood between the full and each reduced model as a measure of relative importance of the attribute, and convert to a proportion.

$$W_{attribute_i} = \frac{|\ell_{full}| - |\ell_{A_i}|}{\sum_j (|\ell_{full}| - |\ell_{A_j}|)}$$

with A_i the reduced model excluding attribute i and j the number of attributes.

3. Calculation of relative preference weights using the coefficient range method:
 - a. For each attribute, calculate the range between the coefficients of the individual levels.
 - b. Convert to a proportion

$$W_{attribute_i} = \frac{\max C_i - \min C_i}{\sum_j (\max C_j - \min C_j)}$$

with C_i the coefficients of the individual levels of attribute i and j the number of attributes.

Although age was included as an attribute in the choice sets of the Therapeutic need domain and in the model, already from the start we did not consider age as a separate criterion for assessing Therapeutic need, but rather as a piece of information respondents need to assess the relative importance of the disease specific life expectancy. We assumed that the utility loss of life year losses – and hence the perceived therapeutic need – might vary with patient age. Therefore, the age of the patient population had to be specified. For the model development, we included age and life expectancy as two separate variables. The relative weights for Therapeutic need were calculated in two different ways:

- Including age in the calculation in order to show the weights corresponding to the way the choice sets were presented to respondents.
- Excluding age in the calculation because in the MCDA application age is not considered for evaluating Therapeutic need.

The relative weights of different criteria were calculated for the entire sample, as well as for subgroups of respondents. We defined subgroups by self-reported age category, by self-reported health status, by self-reported uncertainty of responses, and by number of reminders received.



For self-reported age category, we used the levels as presented to the respondents. For self-reported health status, we reduced the original five levels as follows: "very bad", "bad", and "mediocre" were labelled "not in good health", while "good" and "very good" were labelled "in good health". For the uncertainty of the responses, we calculated a certainty index per respondent:

- We reduced "very uncertain" and "uncertain" to "uncertain". Similarly, we reduced "certain" and "very certain" to "certain".
- Per domain, we used the most frequently reported certainty answer to classify a respondent as certain or uncertain for the answers in that domain.

For each subgroup, the algorithm outlined above was used giving a set of weights per subgroup.

Key points

General approach

- The discrete choice experiment technique was chosen to elicit public preferences, based on the advantages and disadvantages of each technique described in the literature review.
- A web- and paper-based survey was performed in a representative sample of the general Belgian public, drawn at random from the National Registry, stratified by age and sex. 20 000 Belgian citizens were invited to participate.
- The questionnaire was developed in Dutch and French, pre-tested and pilot-tested before launch. A test-retest reliability study was performed.
- Participation in the survey was anonymous.
- 24 versions of the questionnaire were developed, differing only in the contents of the choice sets of the discrete choice experiment, in order to cover a sufficiently large range of scenarios.

Structure of the questionnaire

- The questionnaire had three blocks of questions, relating to different aspects of the reimbursement appraisal process: (1) therapeutic need, (2) societal need and (3) added value of the new treatment.
- For therapeutic need, three criteria were defined: discomfort of current treatment, quality of life with current treatment and life expectancy with current treatment.
- For societal need, two criteria were defined: public expenditure per patient with the disease, prevalence of the disease.
- For added value of the new treatment five criteria were defined: improvement in life expectancy compared to current treatment, improvement in quality of life compared to current treatment, improvement in treatment comfort, reduction in public expenditure of the disease per patient and reduction in prevalence of the disease.
- Respondents had to answer three choice sets for therapeutic need, one for societal need and five for added value.
- In addition, a set of questions relating to demographic and other respondent characteristics were asked.

Data analysis

- A multinomial logit model was estimated to allow the derivation of the relative preference weights for the criteria in each domain.



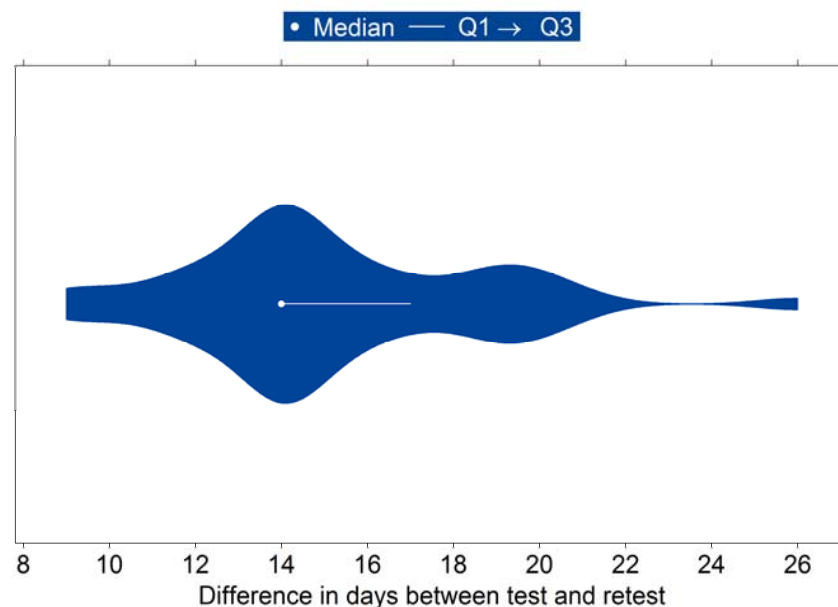
5.2 Test-retest reliability

About half of our test-retest sample were Dutch-speaking women (see Table 12). The median time between test and retest was 14 days (Inter-quartile range (IQR)=3; see Figure 22).

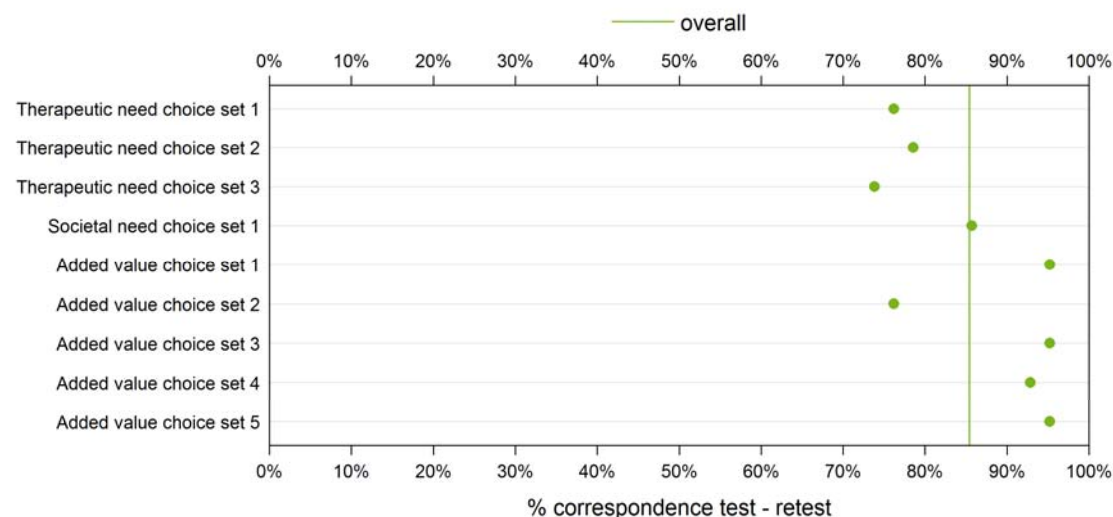
Table 12 – Gender and language distribution of test-retest sample

	French-speaking		Dutch-speaking	
	N	%	N	%
Female	12	28.6%	19	45.2%
Male	1	2.4%	10	23.8%

Figure 22 – Distribution of time between test and retest



The overall agreement between test and retest was Cohen's Kappa = 0.7 (approx. 95% CI: 0.62–0.77). Over all choice sets, the majority of the respondents chose the same alternative in test and retest. In 323 out of the 378 choice sets (85.4%) completed by all respondents in the test phase, the choice was the same in the retest phase. However, this correspondence varied per domain and per question (see Figure 23). Correspondence between test and retest answers was generally less for the therapeutic need domain than for the added value domain. For some questions, the correspondence between the test and the re-test was very high (95% of the respondents made the same choice in the test and re-test phase), whereas for some other questions the correspondence was 74% (question 3 in the Therapeutic need domain). It should be noted that the number of people who participated in the test-retest exercise was rather limited (N=42) and drawn from a selected population (mainly people with higher education and in paid work). This may lead to a biased estimate of the test-retest reliability.

**Figure 23 – Correspondence between answers in test and retest**

5.3 Sample characteristics

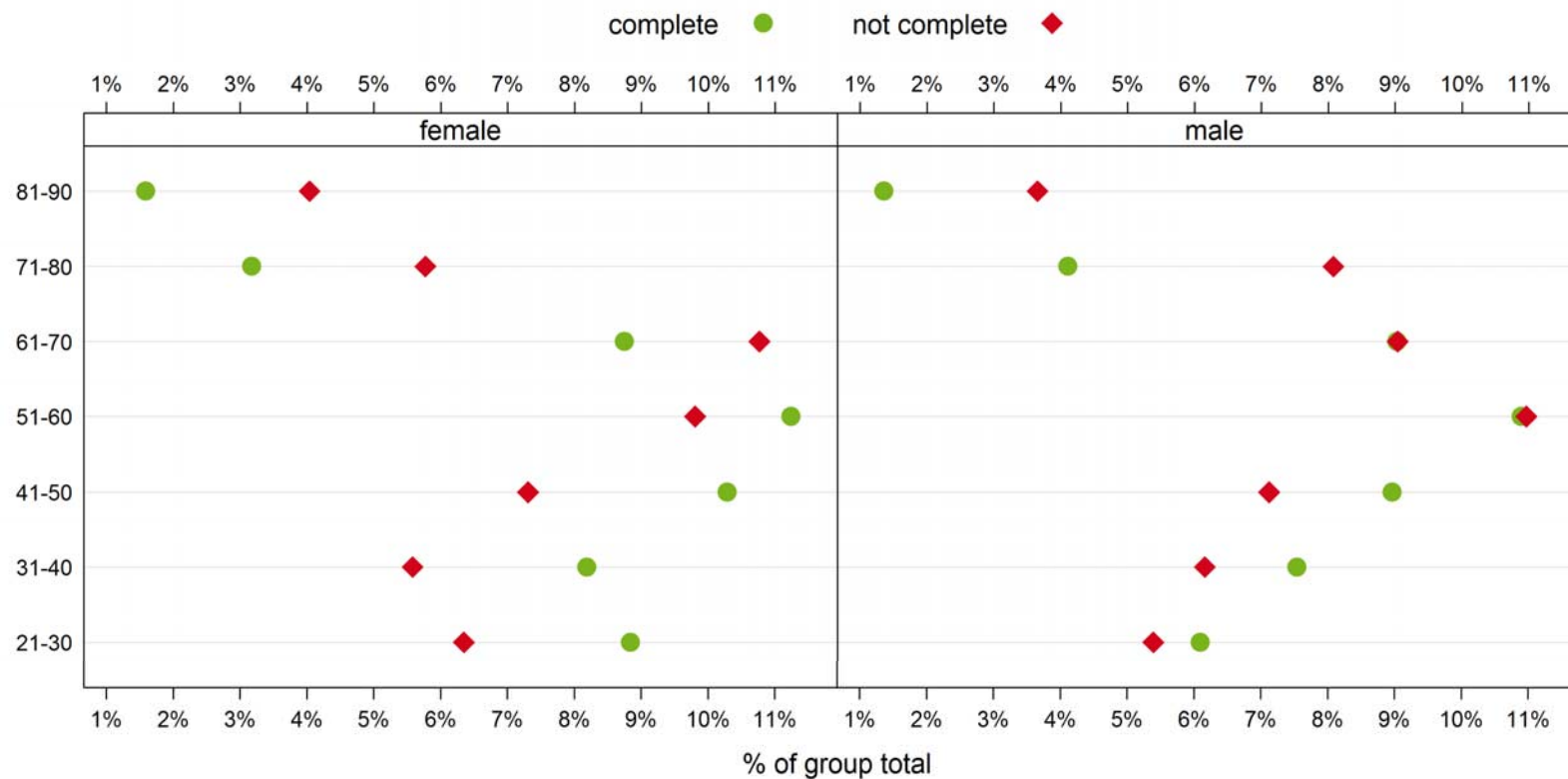
5.3.1 Demographics of the general population sample

Of the 20 000 invited people, 14 had died between the drawing of the sample and the receipt of the invitation letter. Of the remaining 19 986 people invited, 4810 (24%) started the web-survey or sent back a paper questionnaire, and 1027 (5%) explicitly stated that they did not want to participate. Of those who filled out the survey, 27% responded to the initial invitation, while respectively 34%, 19%, and 20% responded after the first to third reminder. Over 91% (3918) completed the survey in a web interface, the others on paper. For comparison, according to the Federal Public Service Economy, about 82% of the Belgian citizens regularly access the internet.³ About 66% of the respondents filled out the Dutch survey and about 34% the French survey.

In total, 4485 respondents (22.4% of invited; 93.2% of respondents) answered all choice sets. Of these, 52.1% were women (compared to 51.3% in the total population between 20 and 89 years of age). Two checks were performed to assess the consistency and possibly the comprehension of the respondent's answers on the choice sets. First, less than 1% of the respondents always chose the first or the second alternative in the nine choice sets. Second, we introduced a dominant choice set halfway through the Added value domain choice sets: all levels of the attributes of the second alternative were "better" than those of the first alternative. About 96% of the general population sample chose the "better" alternative. We excluded 197 (4.4%) respondents from further analysis based on these checks. The analysis sample eventually consisted of 4288 respondents (21.4% of invited; 89.2% of respondents). The comparison of the age and gender distribution between the respondents who did not complete all the choice sets (n=522) and the analysis sample is shown in Figure 24. Tabular data of the graphs in this section can be found in appendix.



Figure 24 – Age and gender distribution of the general population analysis sample (complete) compared to the respondents who didn't complete all choice sets (not complete).





In the analysis sample, the proportions of male and female respondents are comparable to those of the Belgian population ($X^2[1df] = 1.05$, $p = 0.31$) but the proportions of respondents per age category differ from those of the Belgian population ($X^2[6df] = 170$, $p < 0.01$; see Figure 25). For language, the comparison is more difficult. We could not find national data about the mother tongue or registered language of Belgian citizens. The population distribution over regions is not a valid reference, as Dutch-speaking people may live in the French region and vice versa. In our study, respondents chose the language in which they wanted to fill out the survey. Therefore, we used the distribution as shown in a survey from the European Commission from 2005.⁹⁹ It should be emphasized that such a comparison has limitations. First, the data are also only based on a survey and not on official registered population data. Second, evolutions in language distributions since 2005 are not taken into account. Third, we have to make the hypothesis that people chose their mother tongue when filling out our survey. And finally, it does not take into account that respondents with another mother tongue than French or Dutch would have to choose either one of these.

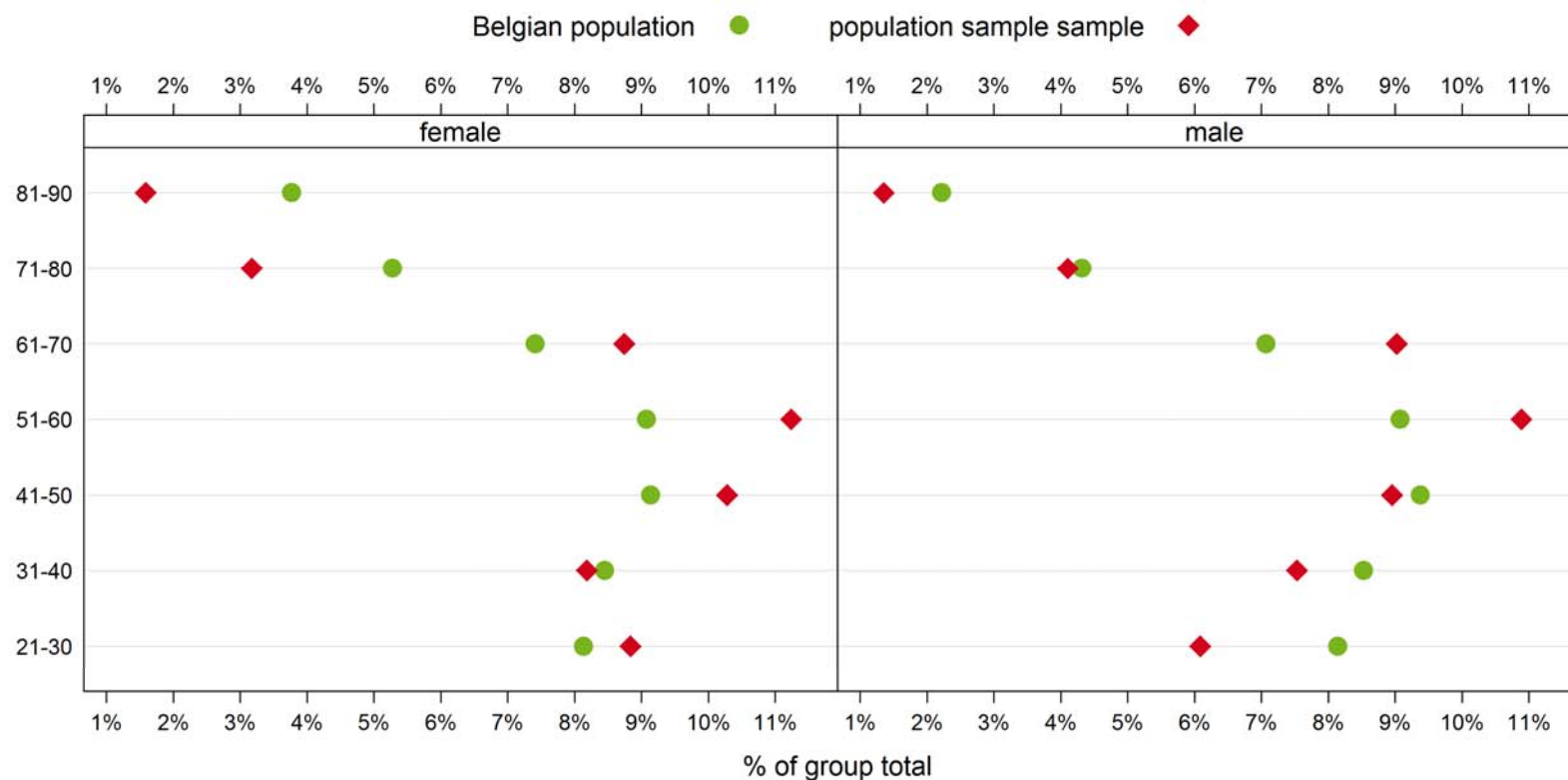
Our sample has more respondents choosing for the Dutch survey version than could be expected from the European Commission survey ($X^2[1df] = 19.1$, $p < 0.01$ (see Table 13).

Table 13 – Language distribution

Language	Sample survey	KCE	Sample survey European Commission
Dutch	62.8%		59.6%
French	37.2%		40.4%



Figure 25 – Age and gender distribution of the general population sample compared to the Belgian population

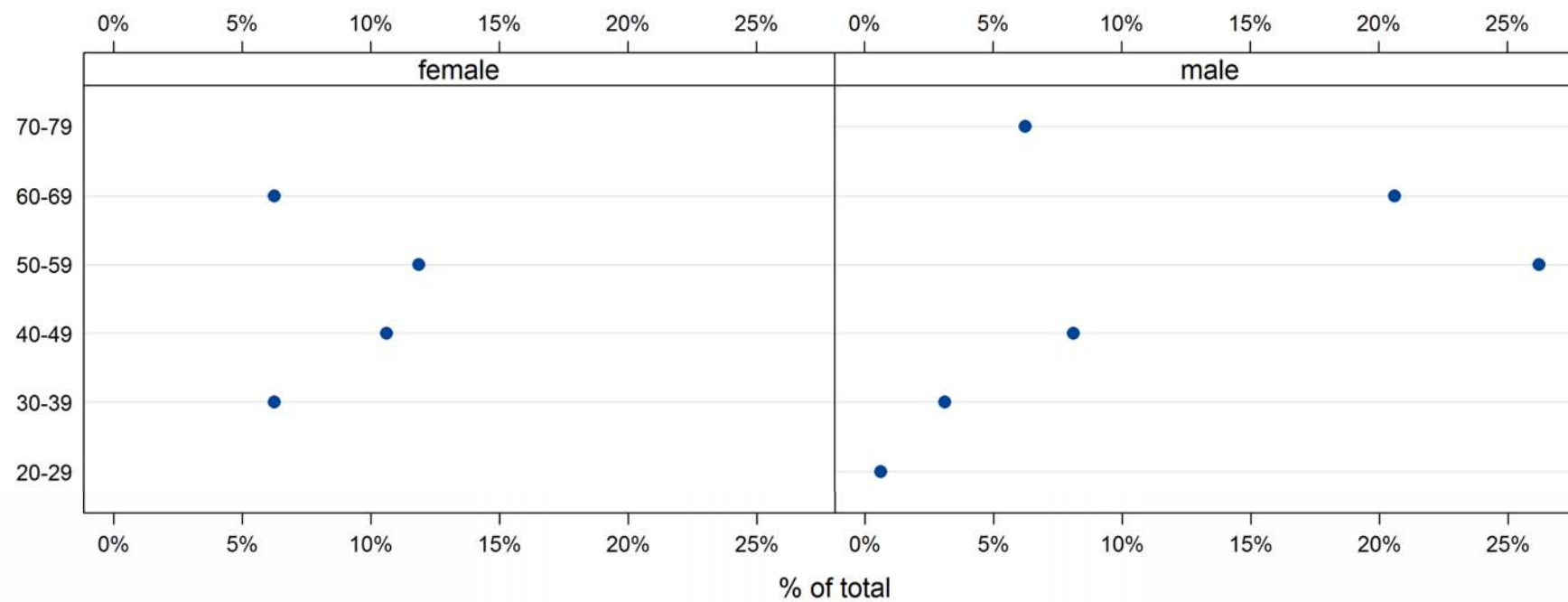


5.3.2 Demographics of the decision makers' sample

Of the 421 invited, 175 (41.6%) responded to the survey. 161 (38.2% of invited, 92% of respondents) answered all choice sets. 60% responded to the initial invitation, while respectively 24%, 12%, and 4% responded after the first to third reminder. All completed the survey in a web interface. About 57% of the respondents filled out the Dutch survey and about 43% filled out the French survey.

We excluded one respondent based on the consistency check described above. The analysis sample then consisted of 160 respondents (38% of invited; 91.4% of respondents).

As we have no age and gender information of the non-responders, we can only show the distribution of the sample (see Figure 26).

**Figure 26 – Age and gender distribution of the decision makers' sample**



The response rates for the invited decision-maker organisations are presented in Table 14. All the advisory committees of the RIZIV – INAMI had a participation rate of more than 45%. The parliamentary committee for health and the senate committee on social affairs had the lowest response rate.

Table 14 – Response rate per decision-maker organisation

Organisation	% complete
College of Medical Directors	61.3%
CTIIMH – CRIDMI	56.3%
Technical Medical Council	53.6%
CTG – CRM	45.3%
FAGG – AFMPS	39.2%
Policy Unit Minister of Public Health and Social Affairs	38.9%
Belgian Advisory Committee on Bioethics	37.2%
Parliamentary Committee for Health	11.6%
Senate Committee on Social Affairs	9.3%

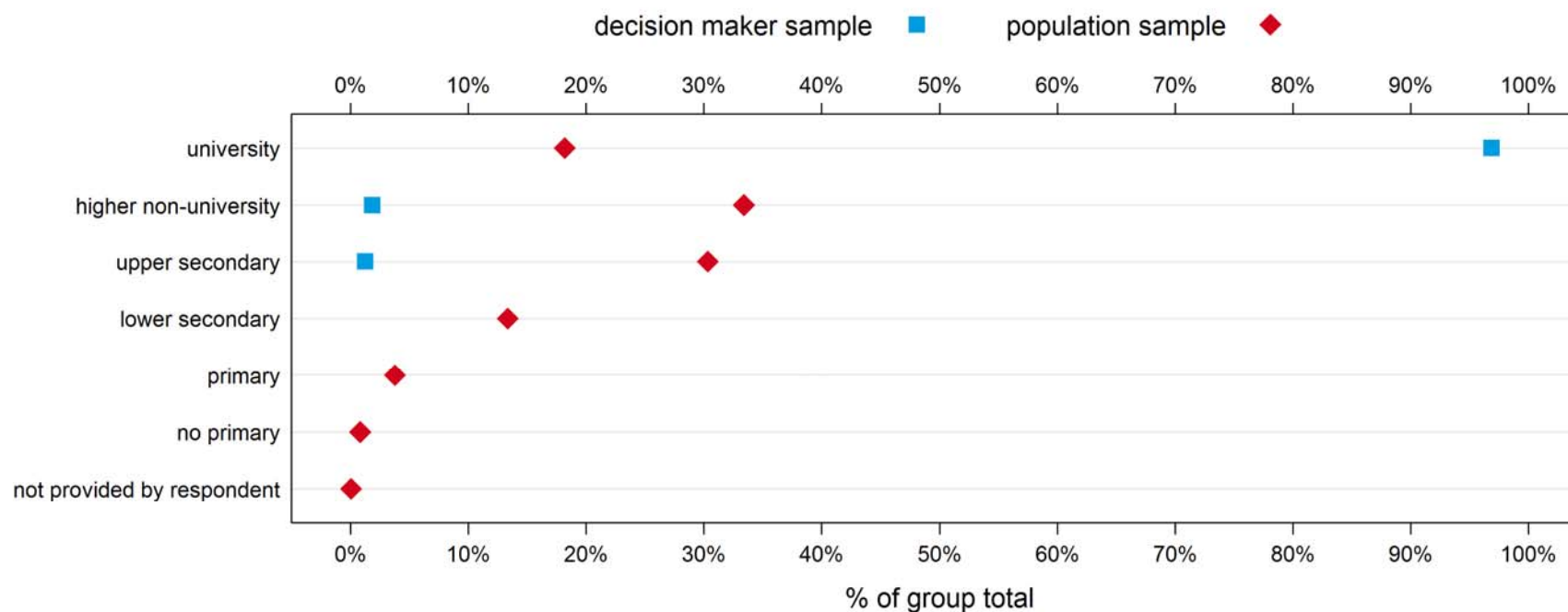
CTIIMH – CRIDMI: Commissie Tegemoetkoming Implantaten en Invasieve Medische Hulpmiddelen / La commission de remboursement des implants et des dispositifs médicaux invasifs

CTG – CRM: Commissie voor Tegemoetkoming Geneesmiddelen / Commission de Remboursement des Médicaments

FAGG – AFMPS: Federaal agentschap voor geneesmiddelen en gezondheidsproducten / Agence fédérale des médicaments et des produits de santé

5.3.3 Comparison of the general population and decision makers' sample

Compared to the general population sample, a much higher proportion of respondents has a university degree (Figure 27). In fact, most respondents in the decision maker sample have a university or higher non-university degree. In the general population sample, the distribution is similarly skewed towards the higher educational groups, be it much less outspoken than in the decision makers group.

**Figure 27 – Distribution of educational levels in the study sample**

Eleven percent of the respondents in the general population sample reported having a serious illness; 32.3% reported to have a relative with a serious illness. In the decision makers' sample, this was 5% and 39.4% respectively. None of the decision makers rated his/her health as bad or very bad. In the general population sample, a small minority rated his health as bad (4.1%) or very bad (0.6%) (Figure 28). These proportions are very similar to those in the Health Interview Survey 2013, an interview survey conducted among 10 000 Belgian citizens composing a representative sample of the Belgian population¹⁰⁰ (see Figure 29).

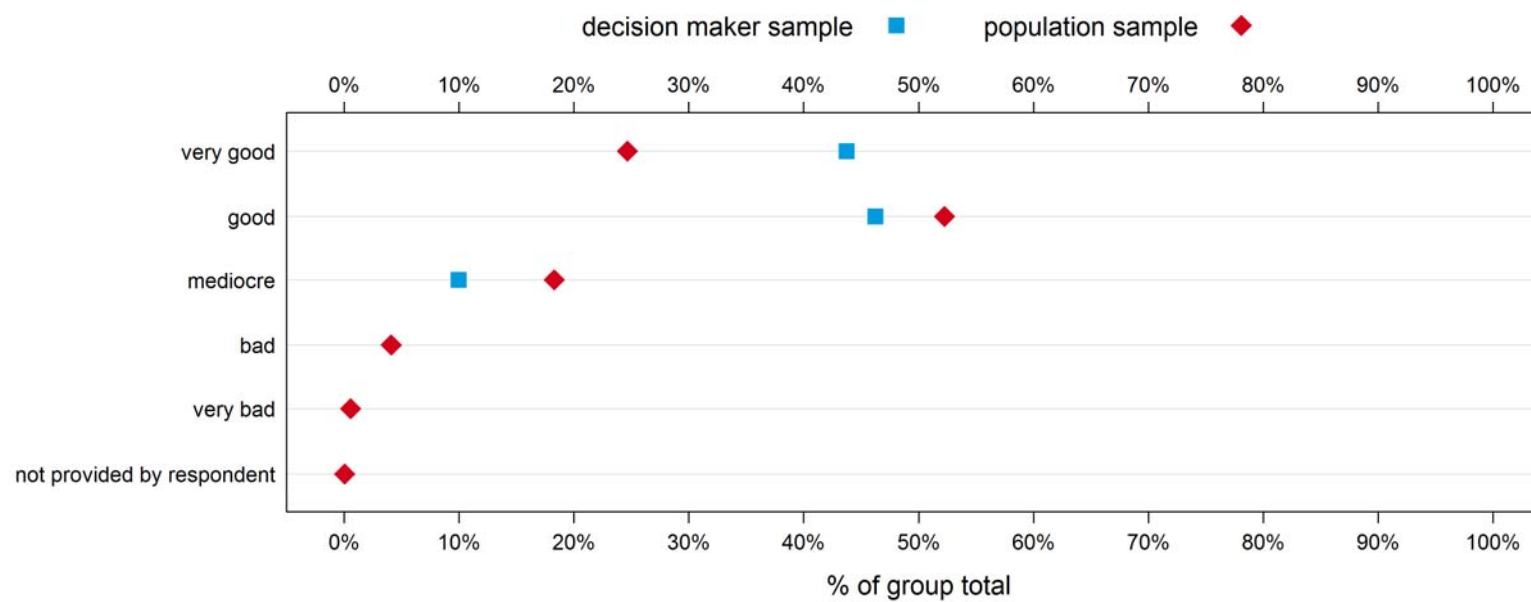
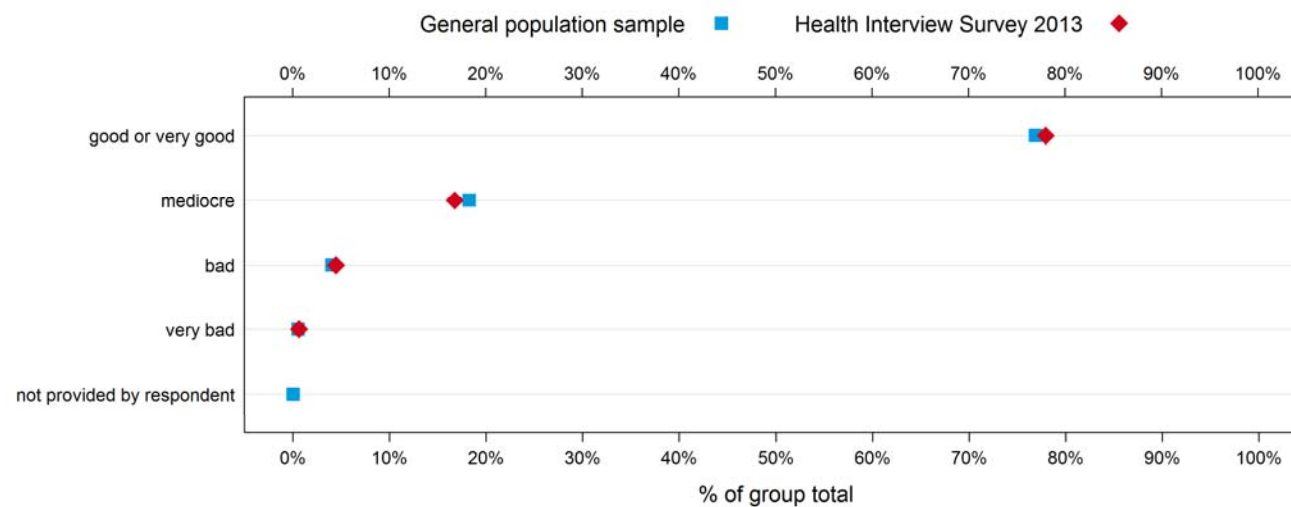
**Figure 28 – Self-reported health status**



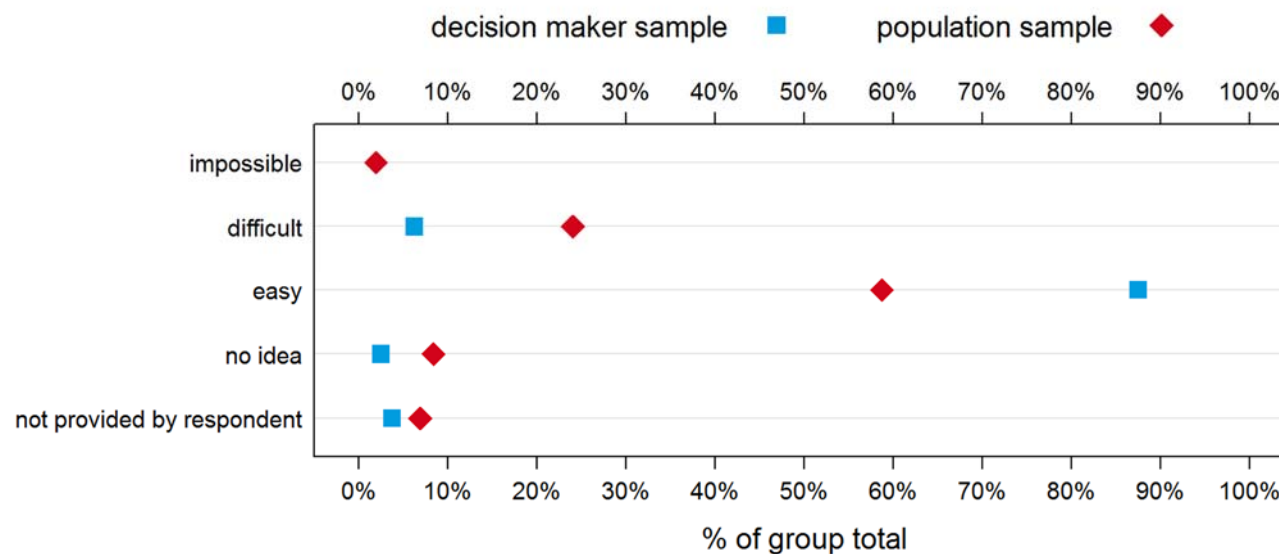
Figure 29 – Self-reported health status in the general population sample, compared to Health Interview Survey 2013





Almost 60% of the respondents in the population sample considered health expenditures to be easily bearable, compared to 87% of the decision makers. More than 20% of the general population sample stated to find health expenditures difficult to bear. Remarkably, about 2.5% of the decision makers answered to have no idea about the affordability of their health care expenditures (Figure 30).

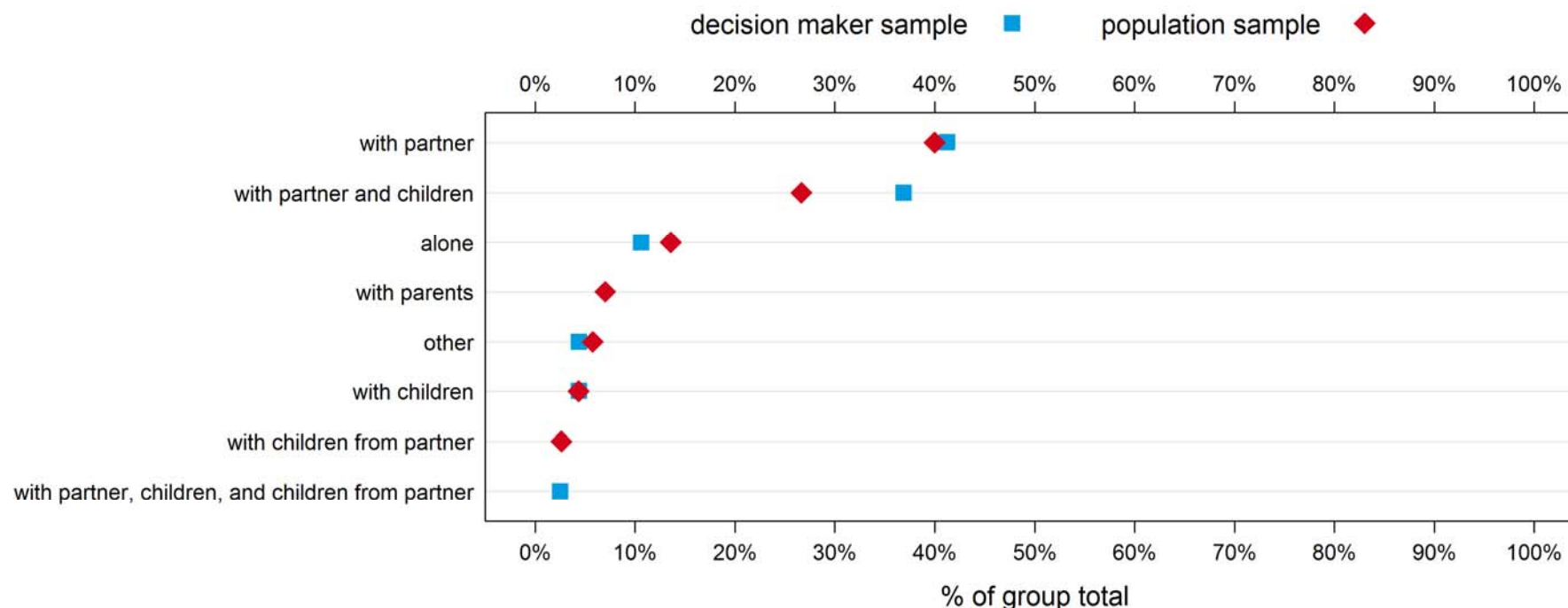
Figure 30 – Affordability of health care



Respondents' living conditions are shown in Figure 31. In both samples, the majority lives with his/her partner or with partner and children.



Figure 31 – Respondents' living conditions



* The category "other" encompasses all categories not included in the graph but available as response options in the survey. Individually, these categories each represented less than 2.5% of the total sample.

Slightly more than 60% of the respondents in the general population sample has a paid activity. 39.2% stated not having a paid activity and 0.2% did not answer this question. The EU Labour Force Survey (EU-LFS, 2013)¹⁰¹ shows an employment rate for Belgium of 67.2% amongst 20 to 64 year olds. Extrapolating to a 20 to 69 years range, assuming no paid activity above the legal retirement age of 65, gives an employment rate of 62%. Given that our question on paid activity resembles quite well the EU-LFS definition of employment, we can compare this 62% to our general

population sample: 67.4% of respondents aged 20 to 69 years old reported to have a paid activity.

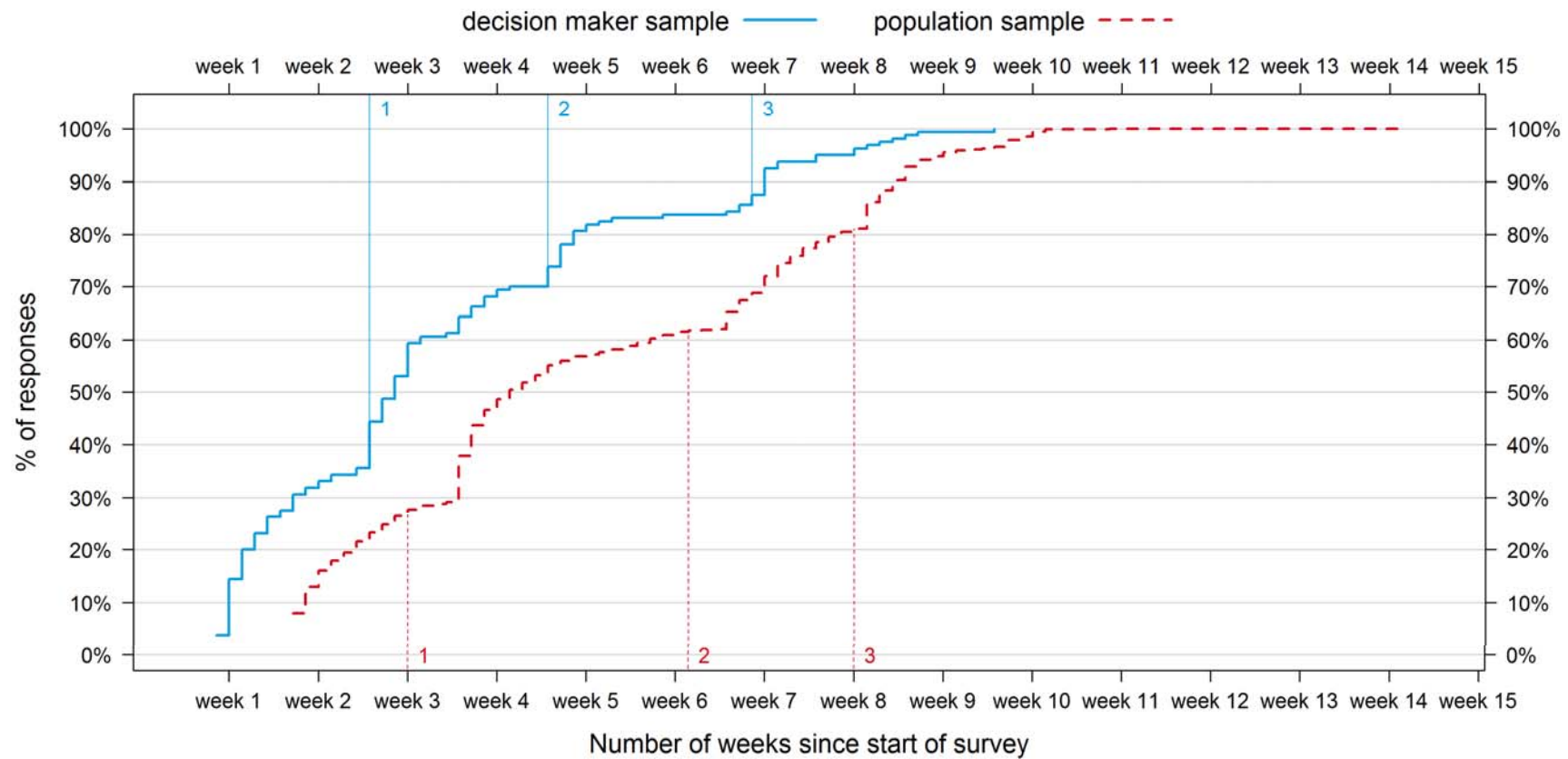
More than 30% were white-collar workers, and almost 21% were retired. Blue-collar workers and civil servants represent both about 10% of the total sample. All other categories are less than 10% in the sample (figure in appendix).



5.3.4 Response by reminder

The effect of sending the reminders is shown in Figure 32 for the paper and electronic version taken together. The figure clearly shows that the reminders evoke an increase in response rate. This applies to both the general population and the decision makers.

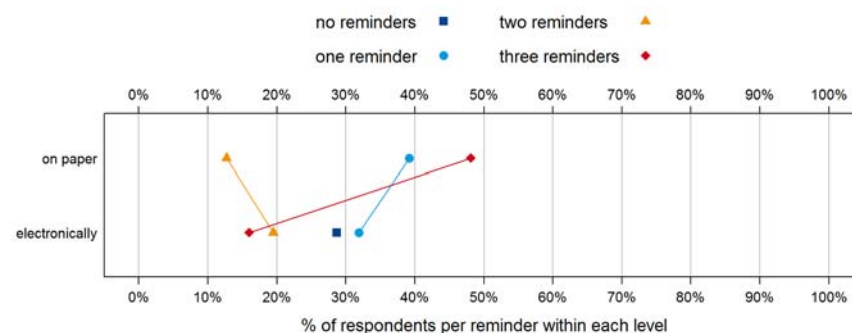
Figure 32 – Reception of responses over time





Almost half of the paper responses were received after three reminders (Figure 33). As decision makers only received the invitation and reminders by e-mail, this figure only relates to the general population sample. No respondents answered on paper immediately after the initial invitation, before the first reminder. This can be explained by the fact that the first reminder was already sent two weeks after the initial invitation. The procedure for requesting a paper version at the National Registry, the National Registry sending out the paper version, completion on paper and sending the questionnaire back might have taken longer than two weeks (10 working days). The observation that almost 50% of the respondents on paper required 3 reminders might also be explained by the fact that the duration of the procedure is long.

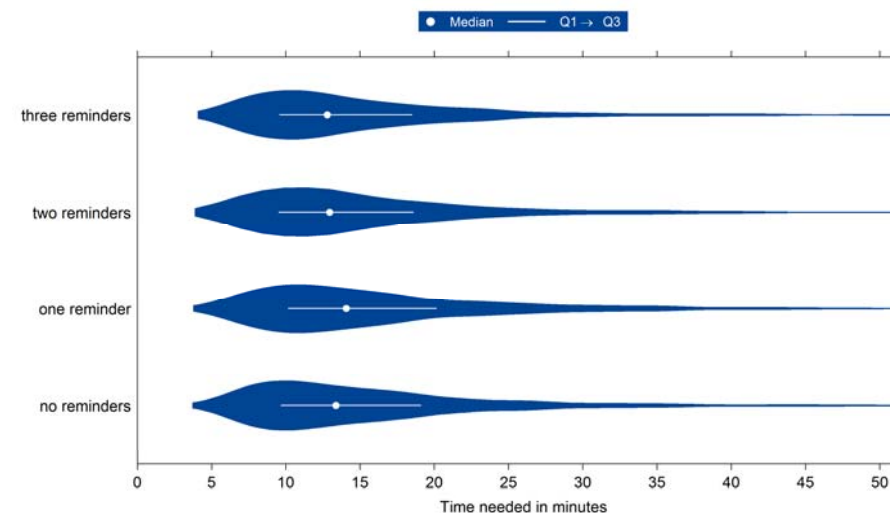
Figure 33 – Response by medium in function of number of reminders



Note: lines provide a visual comparison aid.

There is no difference between the reminder groups in terms of time needed to complete the electronic version (Figure 34).

Figure 34 – Time of completion by number of reminders



Note: 2% of responses are over 53 minutes and are not shown.

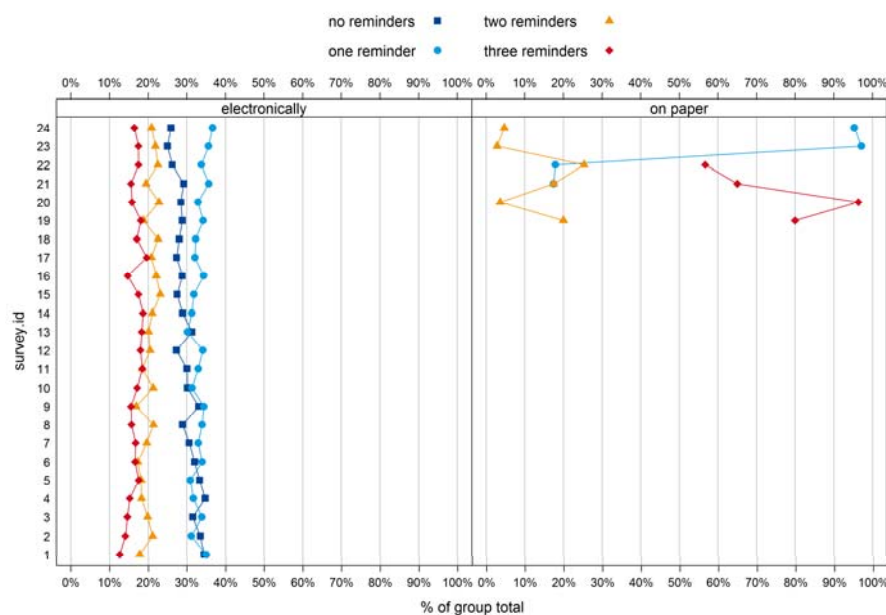
Figure 35 shows the distribution of responses by reminder and by electronic survey version. For each version of the survey, the proportion received after 0, 1, 2 and 3 reminders is presented as a dot. We connected the dots relating to a particular number of reminders each version required. A straight vertical line for e.g. “no reminders” would mean that the proportion of questionnaires filled out after the initial invitation was exactly the same for all versions.

Only versions 19 to 24 were distributed in paper versions. Hence, for these versions, a larger variability in proportions, as compared to the other versions, is observed amongst reminder groups, which is not due to the content of the questionnaires as such.

A general observation is that for all versions, the highest proportion of responses came after one reminder. The smallest proportion of respondents needed three reminders. Differences between versions are limited.



Figure 35 – Proportion of questionnaires returned after initial invitation, one, two or three reminders, by survey version and response medium



5.3.5 Differences in population sample characteristics by reminders

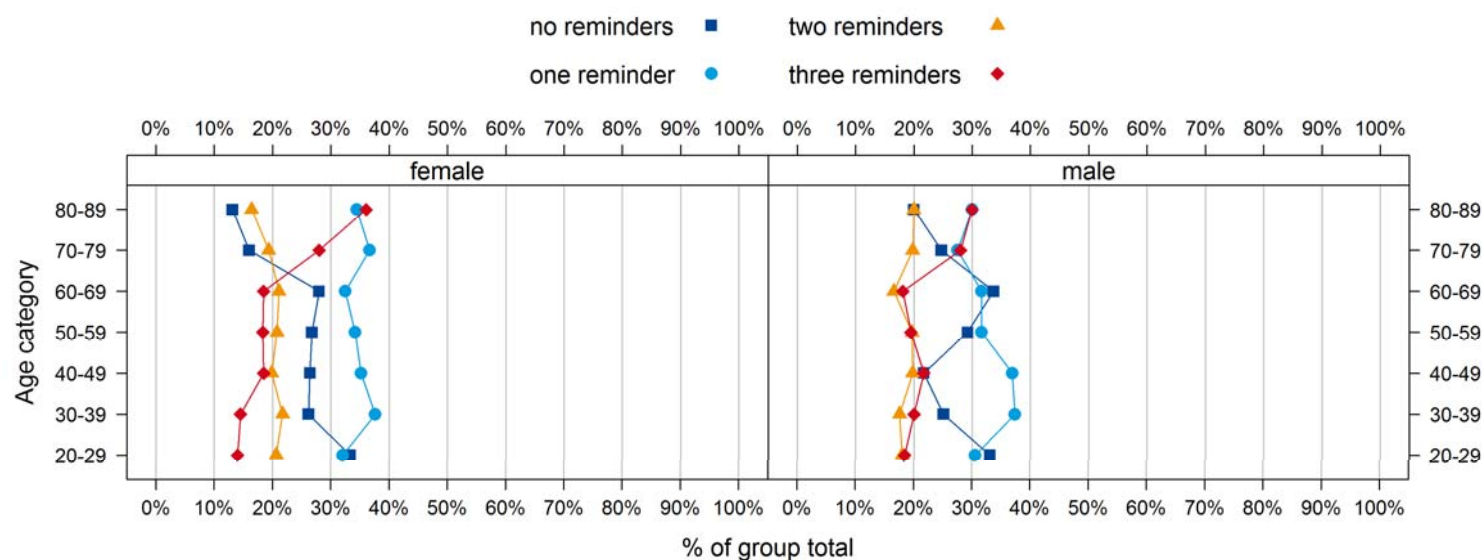
We examined whether the respondents' characteristics differed by number of reminders received. For each variable of interest (e.g. age, education, self-reported health), the proportion of respondents who received no, one, two and three reminders was calculated. These proportions are presented as dots on the figure. We connected the dots relating to one reminder group to create a visual representation of the differences between reminder groups on a specific characteristic. As soon as there are differences in the proportions of reminders across population subgroups, defined by the characteristic of interest, the lines are no longer straight. If the lines for all reminder groups are straight vertical, it means that the characteristic does not correlate with the number of reminders.

If the proportion of respondents with a particular characteristic (e.g. bad self-reported health) requiring 3 reminders is significantly higher than the proportion of respondents with good self-reported health, it may be that people with a bad self-reported health were less easily reached than people with good self-reported health and hence our results might be biased.

Figure 36 shows that the number of reminders differs by age group in both males and females. The proportion of respondents requiring 3 reminders is higher in the older than in the younger age groups.



Figure 36 Number of reminders by age and gender



The number of reminders did not differ by language of the respondent, by them having children or not, or by them having a paid activity or not (see appendix). Also having a serious illness or not, or having a relative with a serious illness or not, did not impact upon the number of reminders required (figures in appendix).

However, there are differences between the number of reminders by family living conditions (Figure 37) as well as by professional status (Figure 38) and educational level (Figure 39). Lower educated people more frequently needed reminders than higher educated people.



Figure 37 – Number of reminders by family living conditions

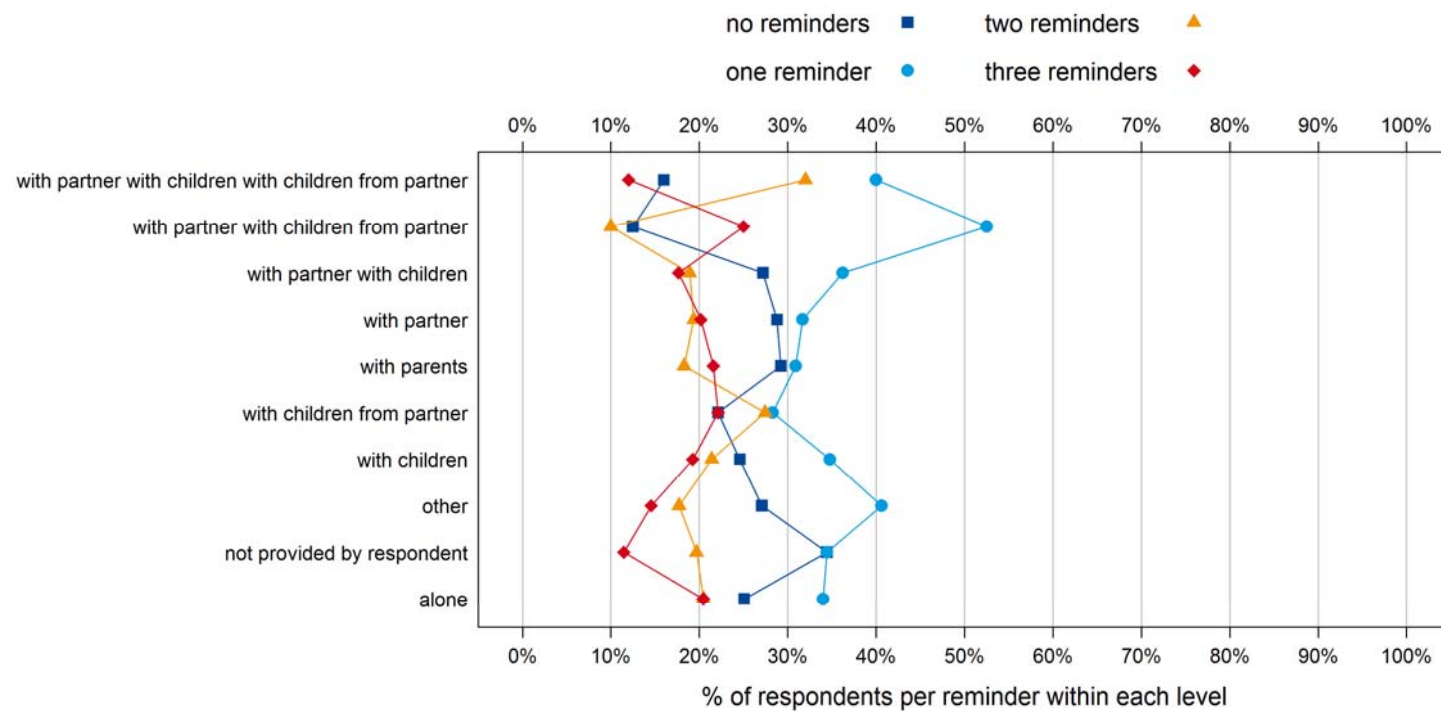
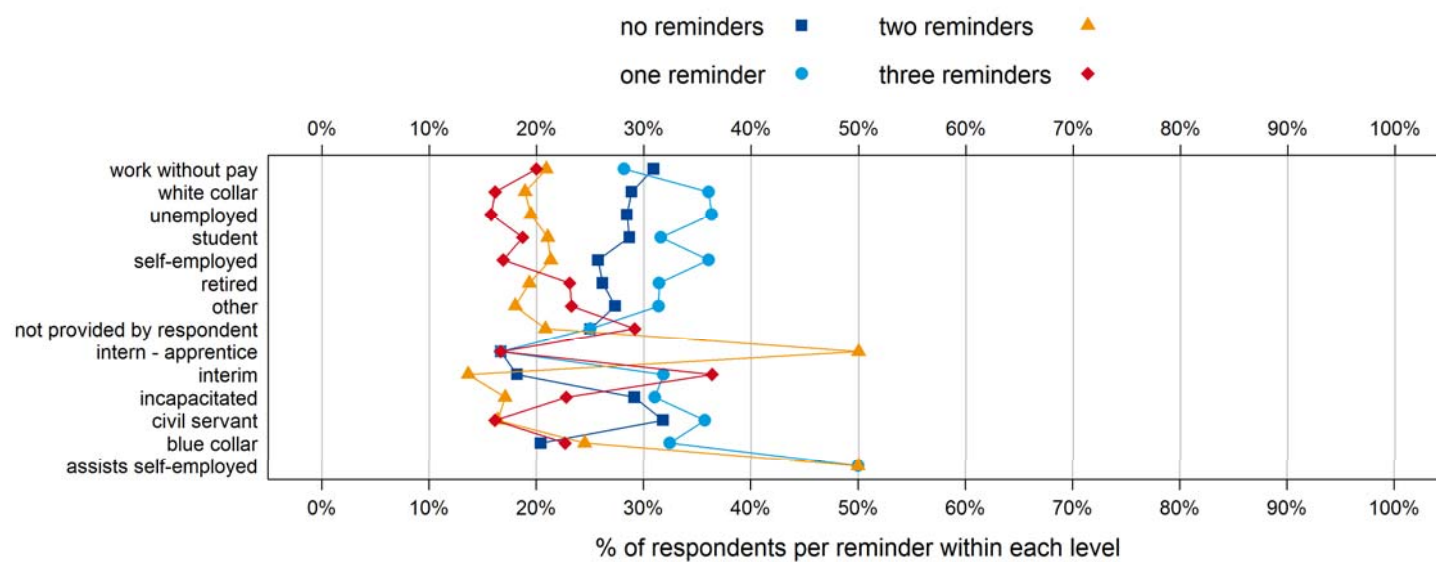
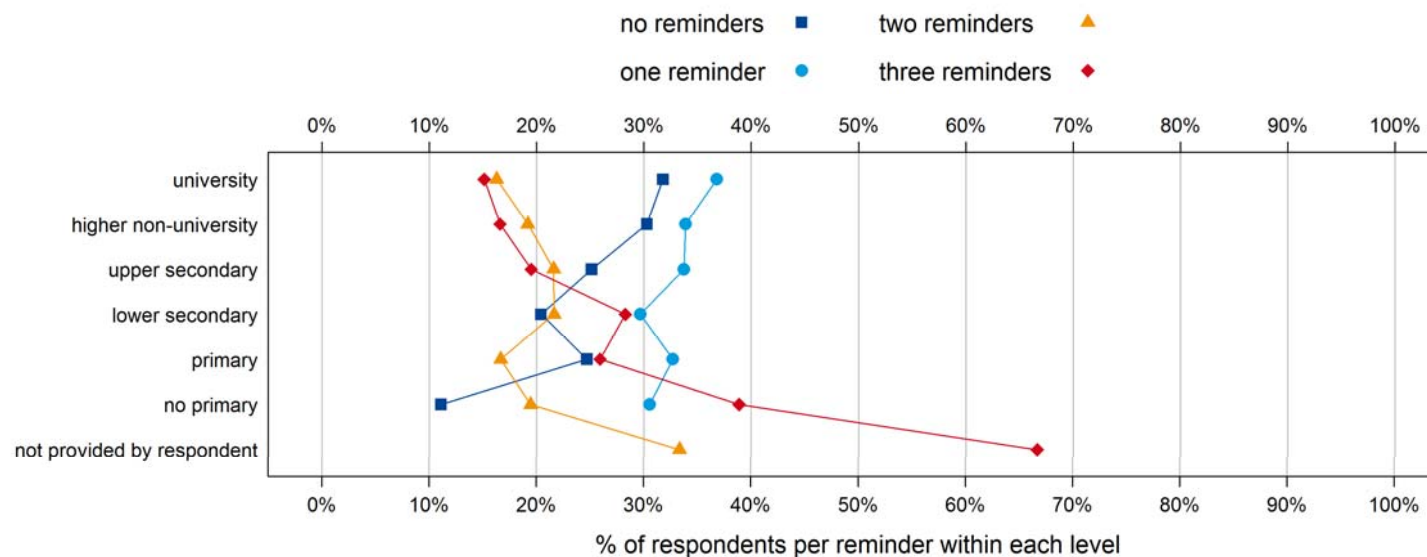


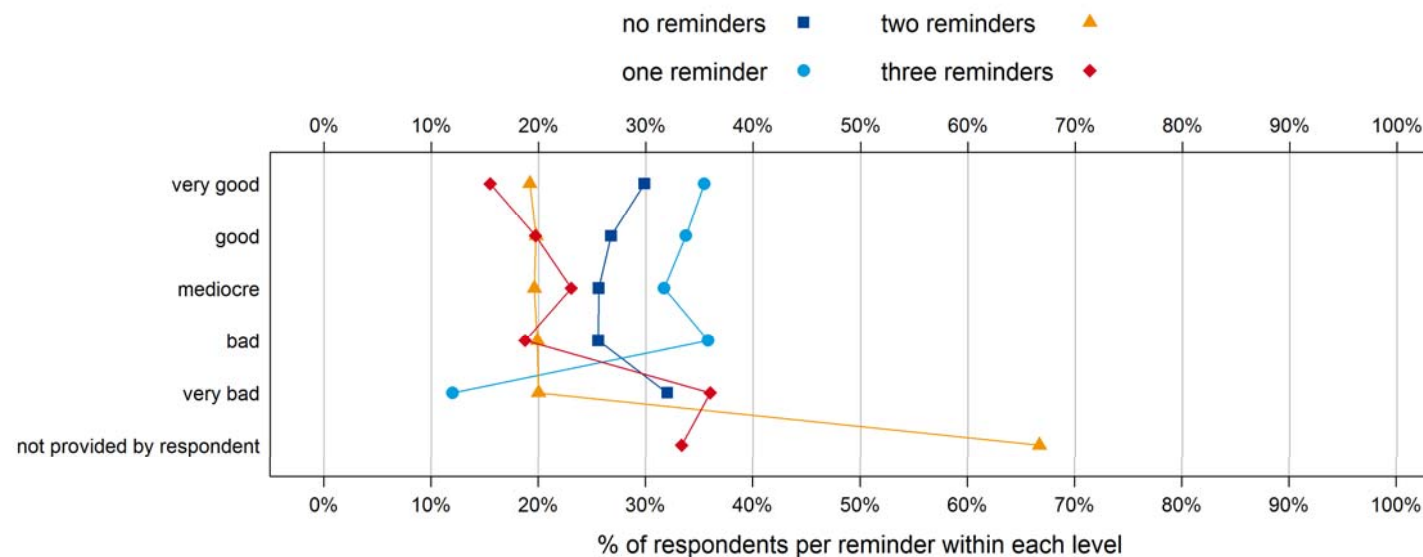


Figure 38 – Number of reminders by professional status



**Figure 39 – Number of reminders by educational level**

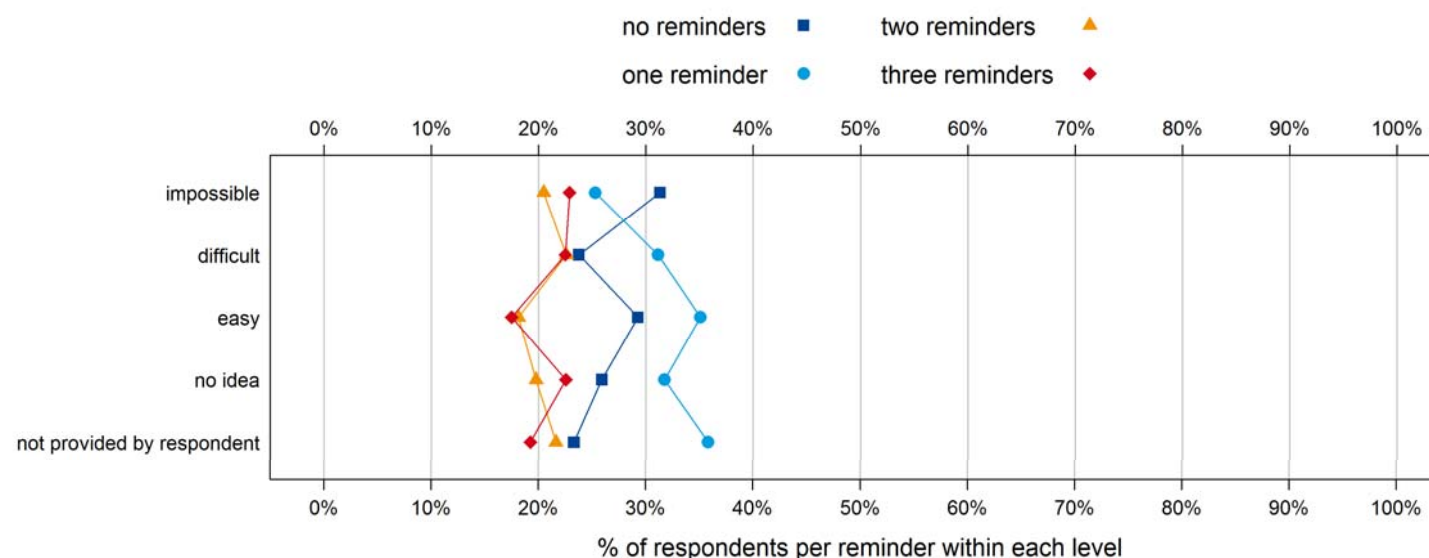
The number of reminders also differs by self-reported current health status (Figure 40). However, except for the respondents in very bad health state, the pattern is not very clear. The proportion of people requiring two reminders in each category of self-reported health is about the same for all self-reported health categories (20%). People in a very bad self-reported health state more frequently required 3 reminders than people with any other self-reported health state. It should be noted, however, that the number of respondents rating their health state as “very bad” is low ($n=25$; 0.6%), which might explain the rather extreme values for this category.

**Figure 40 – Number of reminders by self-reported health status**

The number of reminders also differs by perceived affordability of health care (Figure 41). Again, there is no clear pattern. The proportion of respondents considering health care difficult to afford that answered after two or three reminders was slightly higher than that proportion in the group of respondents considering health care easily affordable, but still the majority of the people in both groups answered after no or one reminder.



Figure 41 – Number of reminders by perceived affordability of health care



5.3.6 Reflections on the sample characteristics

It is unfeasible to verify the representativeness of our population sample on all characteristics that might determine preferences for reimbursement criteria. First, the characteristics determining preferences can only be identified by observing differences in preferences between subgroups. This would mean that a study like ours would have to be done to identify the characteristics on which the sample should be representative. Second, population data are lacking for several characteristics. Therefore, there is no benchmark to which the sample data can be compared.

For the decision makers sample, the differences in response rate between committees should be treated with caution, as some committees have a larger remit than health care decision making and therefore several people in these committees might not be directly involved in health care decision

making (e.g. the senate committee on social affairs and the policy unit of the Minister of Social Affairs). Other organisations do have a close connection with health care but have a different remit than giving reimbursement advice (e.g. the Belgian Advisory Committee on Bioethics). For those committees directly involved in decision making or reimbursement advice, the response rate was relatively high.

Our data show that the decision makers sample is definitely not representative for the general population in terms of demographic characteristics. This observation is less important, however, than the extent to which the preferences of these decision makers represent those of the general public. Given the high response rate, there are strong indications that our sample does reflect the characteristics of the Belgian decision makers in health care.



Our population sample was representative for some demographic and other characteristics (e.g. gender, self-reported health status, paid activity) but not for others (age). The analysis of the data will include sub-group analysis to check whether preferences differ by these characteristics. Especially for the characteristics for which there are doubts about representativeness, this may be relevant.

It was defined in the protocol of the study that three reminders would be sent to increase the response rate. A higher response rate was considered important to increase the precision of the estimated weights. We expected a low response rate, given the complexity of the task. The primary reason for the reminders was thus to increase the response rate.

At the same time, the three reminders allowed to perform subgroup analyses that could provide useful information about the likely bias in our results due to the 77% non- or incomplete response. Late responders were assumed to have preferences closer to those of non-responders.

The clearest conclusion could be drawn from the observed difference between early and late responders in terms of age and educational level. Lower educated elderly people required more reminders than higher educated young people. Sub-group analyses were needed to assess the expected direction of the impact of the underrepresentation of elderly and lower educated people in our sample on the estimated population preferences. In the analysis of the choice sets, we also compare our results with an analysis using a weight to correct the age and gender distribution to correspond to the Belgian population distribution.

Key points

- **The test-retest reliability of the questionnaire can be considered acceptable, with a Cohen's Kappa of 0.7 (0.62-0.77).**
- **Of the 20 000 people invited, 4810 participated in the survey. Responses of 4288 (21.4%) people could be used for the analysis.**
- **The population sample was representative for gender but not for age. Women of 71 or older were underrepresented, as were men below 31 years of age.**
- **For the decision makers, a response rate of 38.2% was achieved. Response rates were highest for committees directly involved in healthcare decision making.**
- **Late responders differed from early responders mainly in terms of age, educational level, where late responders are typically older and lower educated.**

5.4 Choice set analysis: total sample

The results of the models weighted for age and gender distribution are very similar to the unweighted model results. Therefore, we opted to use the unweighted model results. The results of the weighted models are available in appendix.

5.4.1 Reported certainty of choices

For more than three quarters of the choice sets, respondents were certain to very certain about their choice (see Figure 43).



Figure 42 – Reported uncertainty of choices per domain

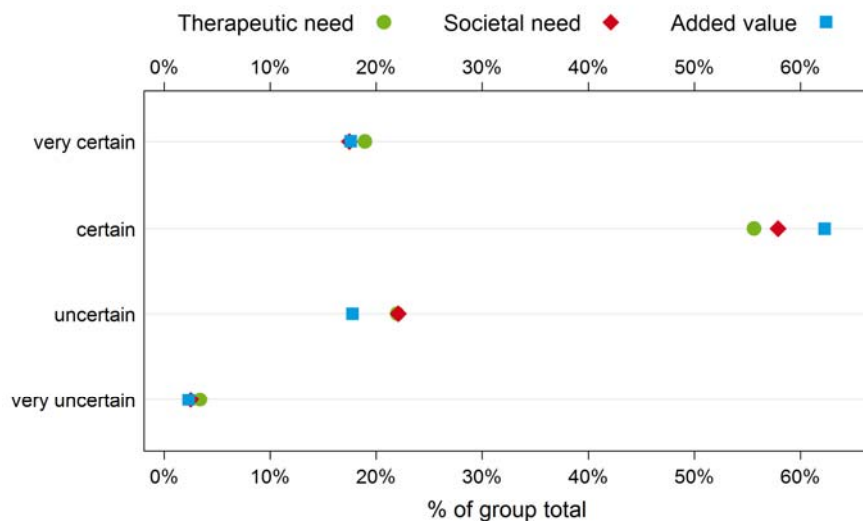
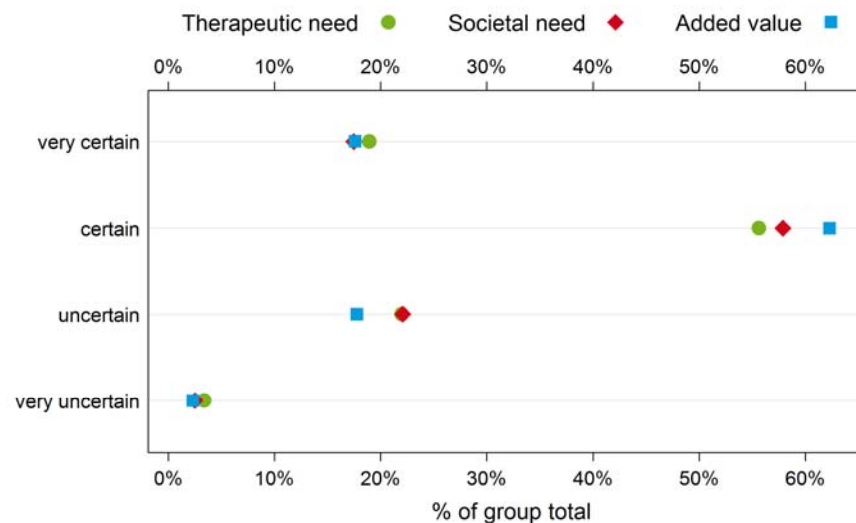


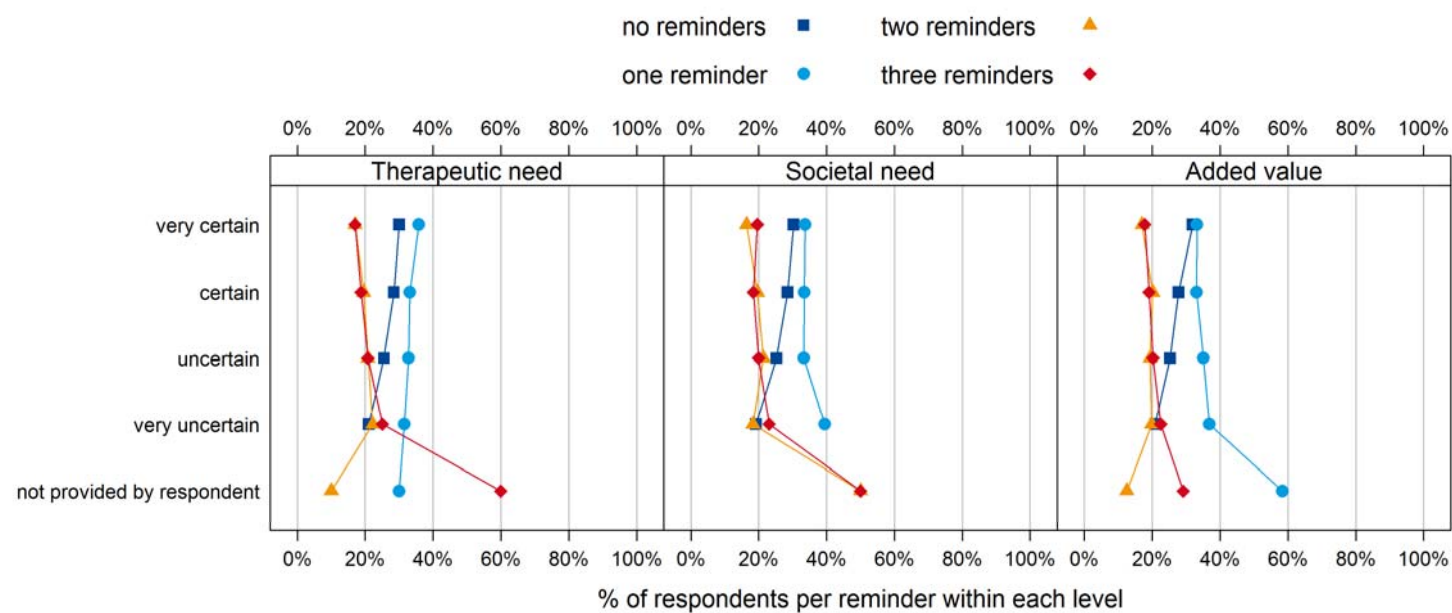
Figure 43 – Reported uncertainty of choices per domain



Choice marked as certain and very certain by respondents are similar across reminder groups (Figure 44). However, respondents needing more reminders tend to be more uncertain of their choices.



Figure 44 – Reported uncertainty of choices per domain and number of reminders





5.4.2 Attribute weights in Therapeutic need domain

5.4.2.1 Predictive value of the model

The estimated model of the general population and decision maker samples predicts fairly well the percentage of choices for each alternative (see Table 15). The percentages refer to the actual, respectively predicted, proportion of respondents who chose the left-side alternative (alternative 1) and the proportion of respondents who chose the right-side alternative (alternative 2). The proportions are not close to 50%. This is because for the Therapeutic need domain, we did not randomize the place (left and right) of the alternatives presented within questionnaire versions across respondents, in order to have a sufficient number of different combinations. Thus, if by coincidence more scenarios with a higher therapeutic need (as judged by the respondents) were on the left-hand side, the percentage “alternative 1” will be higher than 50%.

Table 15 – Actual and predicted percentage of choice for each alternative

	N	Alternative 1		Alternative 2	
		Actual	Predicted	Actual	Predicted
General population	4288	58.4%	60.7%	41.6%	39.3%
Decision makers	160	58.1%	62.1%	41.9%	37.9%

Table 15 and Table 16 suggest a less than perfect fit of the models to the data. For both samples, the model correctly predicts three quarter of the responses for Therapeutic need.

Table 16 – Therapeutic need: goodness of fit statistics

	% of responses correctly predicted by model
General population	75.6%
Decision makers	75.0%

5.4.2.2 General population model

The summary of the full model results for the general population sample are shown in Table 17.



Table 17 – Therapeutic need: model summary for the general population sample

Attribute	Level	Estimated coefficient [°]	Standard Error	t-value	p-value	Significance level
Age	>80y	-1.298	0.029			
	65y - 80y	0.005	0.023	0.237	0.813	
	18y - 64y	0.604	0.029	20.634	<0.001	***
	<18y	0.689	0.029	23.587	<0.001	***
Quality of life given current treatment	8 out of 10	-0.311	0.026			
	5 out of 10	0.063	0.020	3.133	0.002	**
	2 out of 10	0.249	0.019	13.424	<0.001	***
Life expectancy given current treatment	Disease has no impact on life expectancy	-0.188	0.020			
	Patients die 5 years earlier than people without the disease	0.096	0.022	4.279	<0.001	***
	Patients die almost immediately	0.093	0.020	4.5448	<0.001	***
Discomfort of current treatment	little	-0.241	0.019			
	much	0.241	0.014	17.3997	<0.001	***

[°] Results of a multinomial logistic regression model

** significant on the 1% significance level

*** significant on the 0.1% significance level



For a correct interpretation of this model, the following steps need to be taken:

- For each possible scenario, the model needs to be calculated. A scenario is a combination of levels (one for each attribute). Calculating the model for a scenario boils down to adding up the coefficients of the levels of that scenario. As an example, the Therapeutic need for the scenario with quality of life of 2 out of 10, much discomfort of the treatment, in people younger than 18 years old who die almost immediately of the disease is $1.274 = 0.249 + 0.241 + 0.689 + 0.093$ (i.e. the respective coefficient estimates per level). The values thus obtained can be interpreted as utility values or, more appropriate in this context, the level of therapeutic need. A lower value corresponds with a lower probability that the alternative would be chosen as having a higher therapeutic need out of the total set of alternatives, but the values are not probabilities.
- When all model estimates are generated, the scenarios can be listed. The higher the value, the higher the therapeutic need is according to the respondents. The full list for all scenarios that can be described for therapeutic need, with their respective values according to the model, is presented in appendix. The value for therapeutic need has no lower limit nor an upper limit. The value can be negative or positive. Negative values do not mean negative therapeutic need.

Table 18 gives some examples of conditions described according to the included criteria with their respective values of therapeutic need. The higher the value, the higher the therapeutic need is considered by the population. In Figure 45, the probability that a scenario is chosen as having a higher therapeutic need, out of the full set of all possible scenarios, is presented graphically. To reiterate from the methods section, each probability of a particular scenario is calculated as the inverse natural logarithm of the utility value divided by the sum of the inverse natural logarithm of the utility values of all scenarios. For example, the probability for the scenario with quality of life of 2 out of 10, much discomfort of the treatment, in people younger than 18 years old who die almost immediately of the disease is

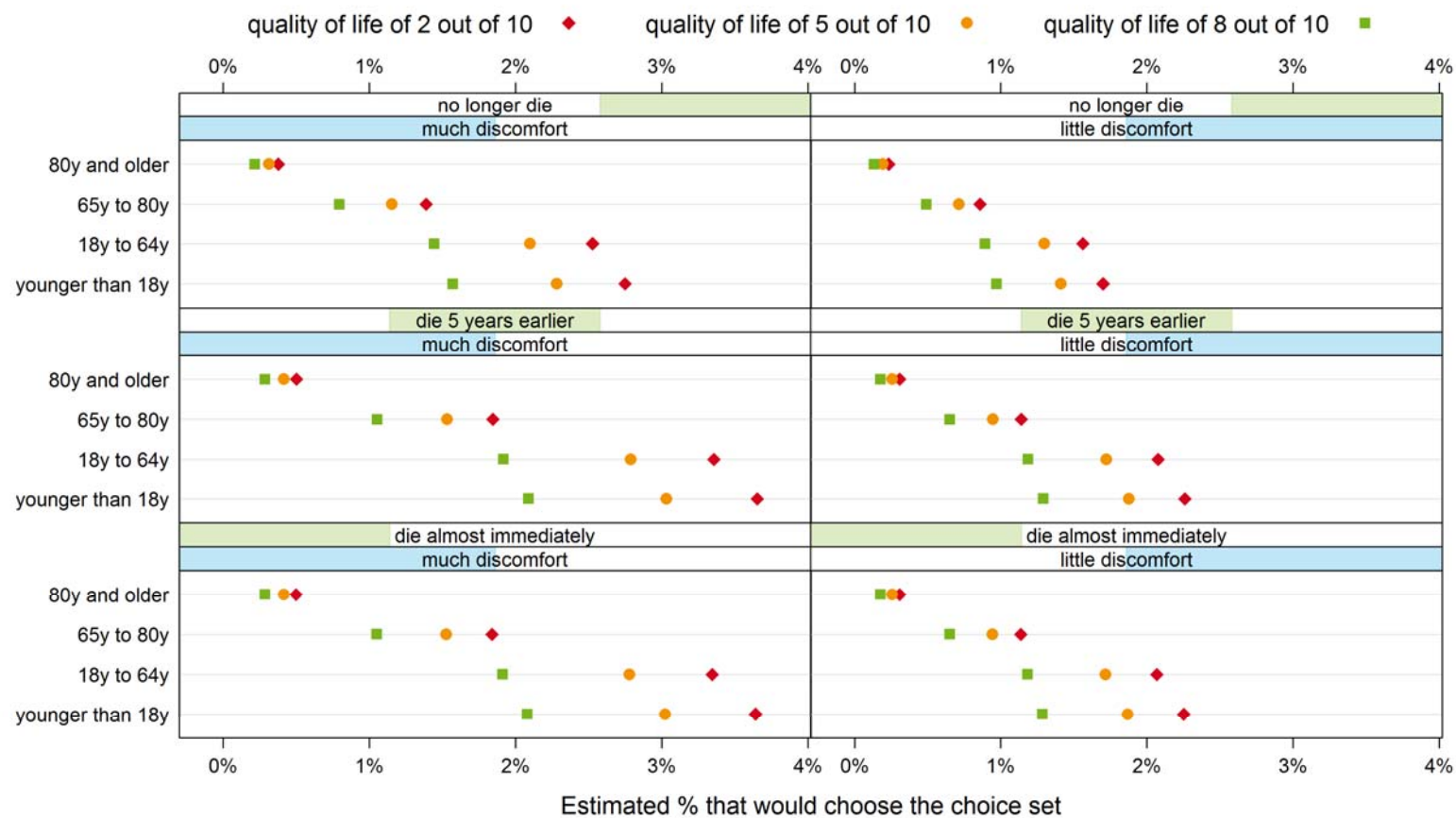
$$0.0365 = \frac{e^{1.274}}{e^{1.274} + e^{1.270} + \dots + e^{-2.038}} = \frac{e^{1.274}}{97.78}$$

In Figure 45, these probabilities are multiplied by 100 to reflect percentages.



Table 18 – Some examples of conditions with their level of therapeutic need according to the model

Quality of life, given current treatment	Discomfort of current treatment	Age	Reduction in life expectancy due to the disease, despite current treatment	Therapeutic need value
2 out of 10	much	younger than 18y	die almost immediately	1.274
5 out of 10	much	18y to 64y	die almost immediately	1.003
2 out of 10	much	younger than 18y	no longer die	1.000
2 out of 10	little	younger than 18y	die almost immediately	0.793
5 out of 10	little	younger than 18y	die 5 years earlier	0.626
5 out of 10	little	18y to 64y	die almost immediately	0.522
5 out of 10	little	younger than 18y	no longer die	0.345
2 out of 10	much	65y to 80y	no longer die	0.322
8 out of 10	little	18y to 64y	die almost immediately	0.148
8 out of 10	little	younger than 18y	no longer die	0.027
5 out of 10	little	65y to 80y	die almost immediately	-0.077
8 out of 10	little	18y to 64y	no longer die	-0.080
8 out of 10	much	65y to 80y	no longer die	-0.175
2 out of 10	much	80y and older	die almost immediately	-0.713
8 out of 10	little	65y to 80y	no longer die	-0.716
5 out of 10	much	80y and older	no longer die	-0.997
2 out of 10	little	80y and older	no longer die	-1.384
8 out of 10	little	80y and older	no longer die	-1.757

**Figure 45 – Probabilities of choosing a scenario as having a higher therapeutic need out of the full set of scenarios, general population**



A general tendency is that older patients (>80 years) are given a lower therapeutic need. However, this correlation between age and therapeutic need assessment is not absolute. This is illustrated by the therapeutic need values for the following groups (Table 18): patients younger than 18, with good quality of life and little treatment discomfort, whose life expectancy is not impacted by the disease versus patients between 65 and 80 years of age who experience much discomfort from their current treatment, have a low quality of life and will not die prematurely from their disease. The <18 year old patients are considered to have a lower therapeutic need (0.027) than the 65-80 year old patients (0.322). Patients in the youngest age groups (<18 and between 18 and 64) and patients between 65 and 80 years of age were assumed to have a similar therapeutic need, if the 65-80 year olds had a very low quality of life (2/10) and the <18-64 year olds had a good quality of life (8/10), all else equal.

The results of the model show that people consider the therapeutic need lowest in patients older than 80 years of age whose remaining life expectancy is not reduced by the disease, have little discomfort from their current treatment and have a good quality of life (-1.757). Patients of 18 years or younger, who are about to die from their disease and who have a very low quality of life with a current treatment that gives a lot of discomfort are considered to have the highest therapeutic need (1.274).

The sign of the coefficient for a particular attribute indicates whether that level of the attribute increases (positive coefficient) or decreases (negative coefficient) the judged therapeutic need in a patient population with particular characteristics (see Table 17). For example, the coefficient for quality of life 2/10 is positive, meaning that citizens judge a population with a quality of life of 2/10 as having a higher therapeutic need than patients with a higher level of quality of life, all else equal.

However, it is incorrect to compare the coefficient of one level of an attribute with the coefficient of one level of another attribute. For example, it is wrong to conclude from the face value of the coefficients that “much discomfort of current treatment (coefficient 0.241)” has almost the same impact on therapeutic need as “having a quality of life of 2/10 with current treatment (coefficient 0.249)”. It is also incorrect to conclude that dying 5 years earlier due to a disease (coefficient 0.096) has an impact on therapeutic need that is about 2.5 times higher than having much discomfort from current treatment (coefficient 0.241). Instead, *differences* in coefficient values must

be compared to make meaningful statements. For example: a change from quality of life of 2/8 to quality of life of 5/8 (difference in coefficients = 0.186) has an impact on therapeutic need that is about 1.5 times higher than a change from “patients die almost immediately from the disease” to “patients do not die from the disease” (difference in coefficients = 0.281).

Exploration of the ranges of coefficients between attribute levels, and also of the model estimates for each scenario (see appendix), offers interesting additional insights (data shown in Table 19):

- People do not discriminate very clearly between the very young (<18 years of age) and the working age adults (18-64 years of age) when judging therapeutic need, i.e. compared to patients between 65 and 80 years, the therapeutic need in patients aged <18 years, all else equal, is 0.68 higher and in patients between 18 and 65 years it is 0.60 higher. Compared to patients between 18-64 years of age, the therapeutic need in patients aged <18 years is only 0.09 higher, all else equal.
- The change in therapeutic need is not proportional to the change in the quality of life score on a 0 to 10 scale. Moving from a quality of life of 8/10 to 5/10 has a relatively higher impact on the value for therapeutic need than a change from 5/10 to 2/10, even though the difference in points on the 0-10 scale is the same for both changes. This means that people value avoiding quality of life loss in patients that currently have a rather good quality of life more than avoiding further loss in quality of life in patients who are already in a bad quality of life state.
- A rather strange finding is that the respondents did not clearly discriminate between “die almost immediately” and “die 5 years earlier”; i.e. the therapeutic need value for a scenario in which patients “die almost immediately” is equal to the therapeutic need value of a scenario in which patients “die 5 years earlier”, all else equal. It might suggest that in their responses people dichotomized this attribute into “lethal disease” and “non-lethal disease”, where both levels indicating premature death are included in “lethal disease”.
- The data also demonstrate the trade-off people make between quality of life and life expectancy for the judgment of therapeutic need. For example, a population with a quality of life of 5/10 that suffers from a non-lethal disease has a similar therapeutic need as a population with a quality of life of 8/10 that dies 5 years earlier from its disease



(scenarios compared are highlighted in the table in appendix). In other words: a lower quality of life and a low impact on life expectancy are equivalent to a better quality of life and a higher impact on life expectancy.

Table 19 – Differences in coefficients between attribute levels

Age (years)	>80	65-80	18-64	<18
>80	0			
65-80	1.3*	0		
18-64	1.90	0.60	0	
<18	1.99	0.68	0.09	0

* This figure represents the difference between the coefficient of the level “65-80 years of age” and the coefficient of the level “>80 years of age” in the attribute “age”. It means that therapeutic need is considered 1.3 points higher in 65-80 year olds than in >80 year olds, all else equal. The figure has no unit but reflects a change in “value” of need. The absolute value of the figures in this table can be interpreted relative to each other and also relative to the figures in the following tables.

QoL	8/10	5/10	2/10
8/10	0		
5/10	0.37	0	
2/10	0.56	0.19	0

^d Some decision makers informed us that, even though they were a member of the institution, they were mainly involved in other policy issues than

Life expectancy	No longer die	Die 5 years earlier	Die almost immediately
No longer die	0		
Die 5 years earlier	0.29	0	
Die almost immediately	0.28	0.01	0

Discomfort	Little	Much
Little	0	
Much	0.48	0

5.4.2.3 Decision makers model

Results of the model for the decision maker sample are shown in Table 20. Comparisons between the results of the general public and the decision makers should be treated with caution, as the sample of decision makers is rather small and therefore the observations for this group more uncertain (i.e. confidence intervals around the estimated coefficients and weights are larger). Nevertheless, it should also be kept in mind that this issue could only have been solved by increasing the number of choice questions for the decision makers. The population of decision makers is small and the response rate already high, hence increasing the number of respondents is unlikely to be feasible^d. Increasing the number of choice questions, however, might reduce the response efficiency.

healthcare reimbursement (e.g. food safety, registration of medical products, bio-ethics)



Table 20 – Therapeutic need: model summary for the decision maker sample

Attribute	Level	Estimated coefficient ^o	Standard Error	t-value	P-value	Significance level
Age	>80y	-1.290	0.156			
	65y - 80y	-0.004	0.118	-0.031	0.975	
	18y - 64y	0.760	0.169	4.506	<0.001	***
	<18y	0.534	0.152	3.507	0.001	***
Quality of life	8 out of 10	-0.469	0.140			
	5 out of 10	0.095	0.103	0.923	0.356	
	2 out of 10	0.374	0.098	3.840	<0.001	***
Life expectancy	no longer die	-0.373	0.114			
	die 5 years earlier	0.115	0.119	0.965	0.334	
	die almost immediately	0.258	0.107	2.398	0.017	*
Discomfort	little	-0.191	0.095			
	much	0.191	0.071	2.698	0.007	**

^o Results of a multinomial logistic regression model

* significant on the 5% significance level

** significant on the 1% significance level

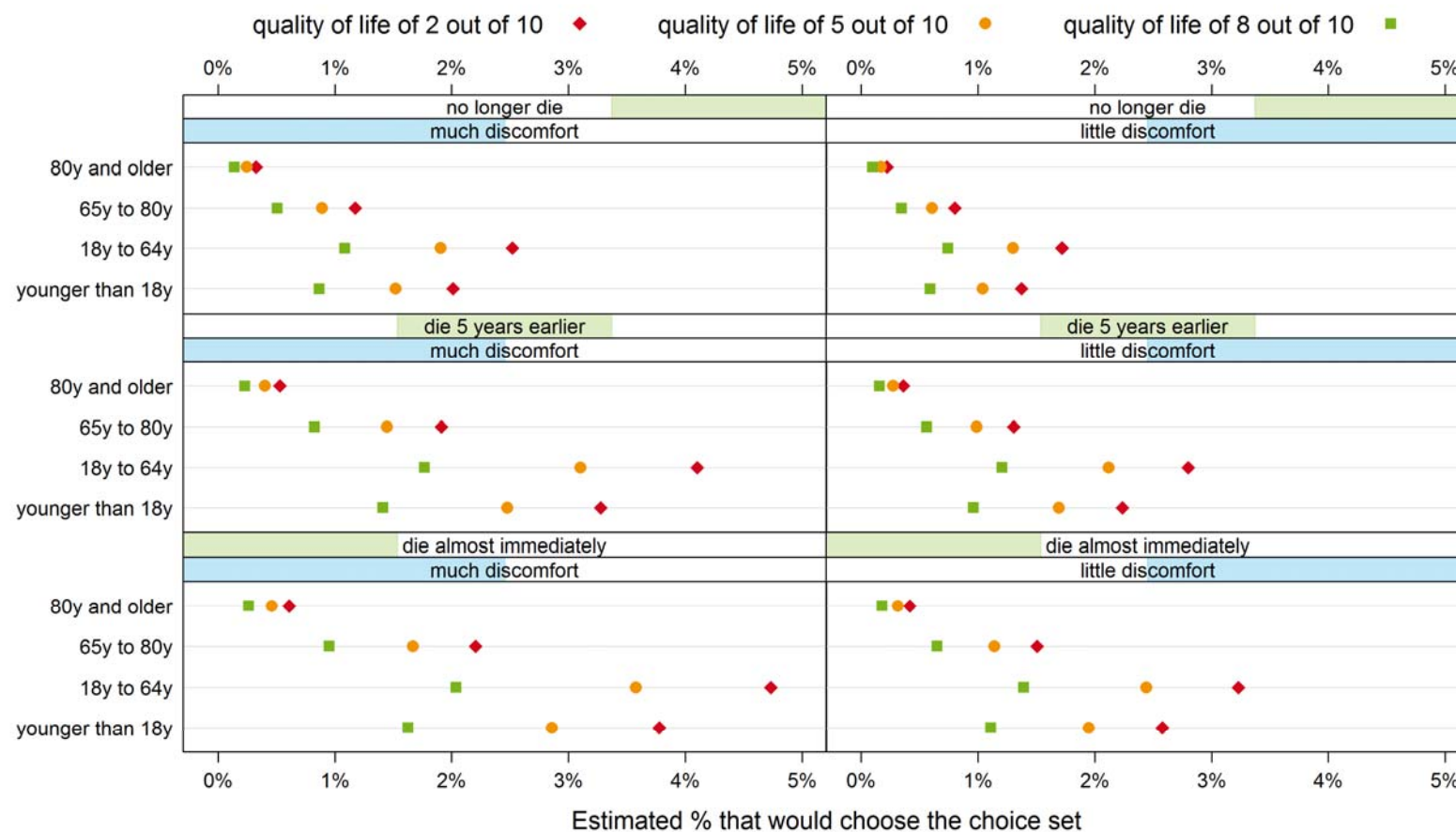
*** significant on the 0.1% significance level

The probability that decision makers will choose a scenario as having a higher therapeutic need out of the full set of all possible scenarios, are presented in Figure 46. A similar pattern as for the general public is observed, except for the fact that the youngest age group (<18 years of age)

does not systematically get higher priority than the 18 to 64 year age group. On the contrary, decision makers frequently judged the therapeutic need to be higher in the 18-64-year olds than in the less than 18-year olds, *ceteris paribus*.



Figure 46 – Probabilities of choosing a scenario as having a higher therapeutic need out of the full set of scenarios, decision makers



5.4.2.4 Weights for the therapeutic need criteria

The log-likelihoods and log-likelihood differences, used for the calculation of the relative weights of the different attributes by the log-likelihood method, are shown in Table 21 for the general population and in Table 22 for the decision makers (see section 5.1.13 for details on the different methods)

**Table 21 – Log-likelihood of models in the Therapeutic need domain, general population sample**

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes	Likelihood ratio test
Full model	-6618			
Model without age	-8637	-2019.4	0.847	$X^2[df=5] = 4038.7$ ($p < 0.01$)
Model without discomfort	-6775	-157.2	0.066	$X^2[df=7] = 314.3$ ($p < 0.01$)
Model without quality of life	-6775	-157.1	0.066	$X^2[df=6] = 314.3$ ($p < 0.01$)
Model without life expectancy	-6668	-50.2	0.021	$X^2[df=6] = 100.5$ ($p < 0.01$)

Table 22 – Log-likelihood of models in the Therapeutic need domain, decision maker sample

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes	Likelihood ratio test
Full model	-244.2			
Model without age	-315.4	-71.2	0.74	$X^2[df=5] = 142.3$ ($p < 0.01$)
Model without quality of life	-257.2	-13.0	0.14	$X^2[df=6] = 25.9$ ($p < 0.01$)
Model without life expectancy	-252.1	-7.9	0.08	$X^2[df=6] = 15.9$ ($p < 0.01$)
Model without discomfort	-247.9	-3.7	0.04	$X^2[df=7] = 7.5$ ($p < 0.01$)

The results show that in both samples age is the most important attribute for making choices between scenarios. However, age was included to interpret life expectancy and not as a decision criterion because age is most often not characteristic of a disease but of a particular patient and can therefore not easily be used in an MCDA (see 5.1.13). Therefore, we recalculated the weights without the attribute “age” (Table 23 and Table 24).

By doing so, we distinguished between “disease-specific impact on life expectancy” and “age-specific impact on life expectancy”. “Impact on life

expectancy” was formulated in our survey as: “patients do not die from the disease”, “patients die 5 years earlier from the disease than people who do not have the disease” and “patients do not die from the disease”. Given this formulation, the weight of the attribute “age” is more likely to reflect the impact of “age-specific impact on life expectancy”, whilst the attribute “life expectancy” is more likely to reflect the impact of “disease-specific impact on life expectancy”.

**Table 23 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, general population sample**

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes
Full model	-6618		
Model without discomfort	-6775	-157.2	0.43
Model without quality of life	-6775	-157.1	0.43
Model without life expectancy	-6668	-50.2	0.14

Table 24 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, decision maker sample

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes
Full model	-244.2		
Model without quality of life	-257.2	-13.0	0.53
Model without life expectancy	-252.1	-7.9	0.32
Model without discomfort	-247.9	-3.7	0.15

The general public considers the impact of a disease on quality of life equally important for therapeutic need assessment as the discomfort of current treatment. The impact of a disease on life expectancy has only limited importance for the assessment of therapeutic need.

These results need to be considered in relation to the results in Table 23 and the earlier observation that people seem to dichotomise impact on life expectancy into “lethal” and “non-lethal”. The general preference for giving more weight to the needs of the younger age group (<18-64) may be explained by the higher number of life years lost in younger patients in case of a lethal disease.

The coefficient ranges, used for the calculation of attribute weights by the coefficient range method, are shown in Table 30 for the general population and in Table 31 for the decision makers (see section 5.1.13 for details).

Table 25 – Coefficient range weights in the Therapeutic need domain, general population sample

Model	Coefficient range	Proportion
Age	1.99	0.60
Quality of life	0.56	0.17
Discomfort	0.48	0.15
Life expectancy	0.28	0.09



Table 26 – Coefficient range weights in the Therapeutic need domain, decision maker sample

Model	Coefficient range	Proportion
Age	2.05	0.52
Quality of life	0.84	0.22
Life expectancy	0.63	0.16
Discomfort	0.38	0.10

As for the log-likelihood method, we recalculated the relative preference weights without the attribute “age” (Table 27 and Table 28).

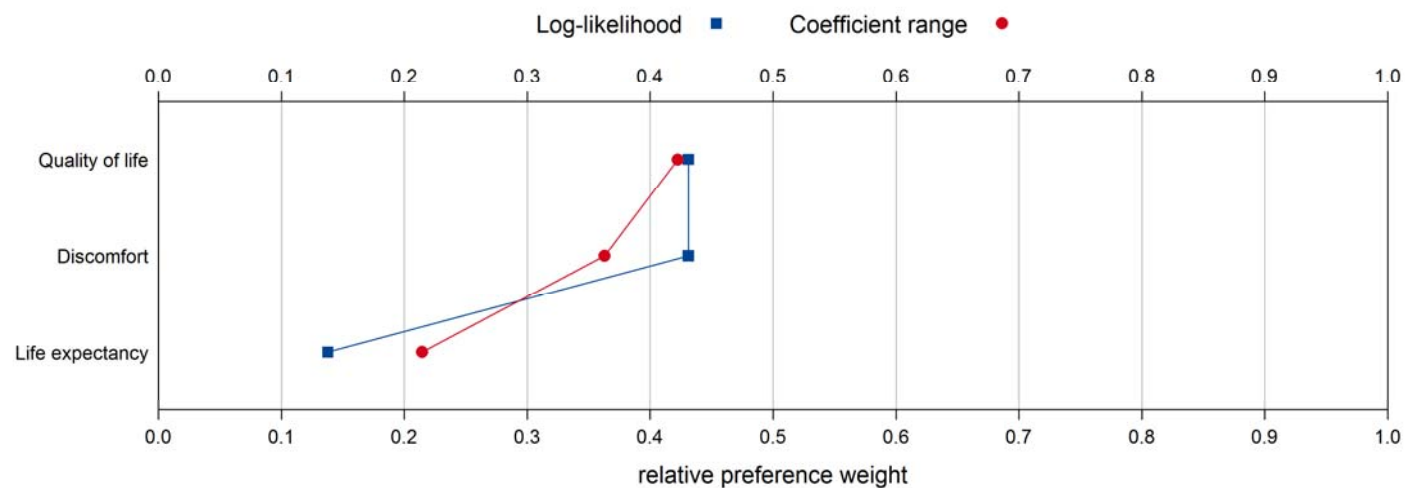
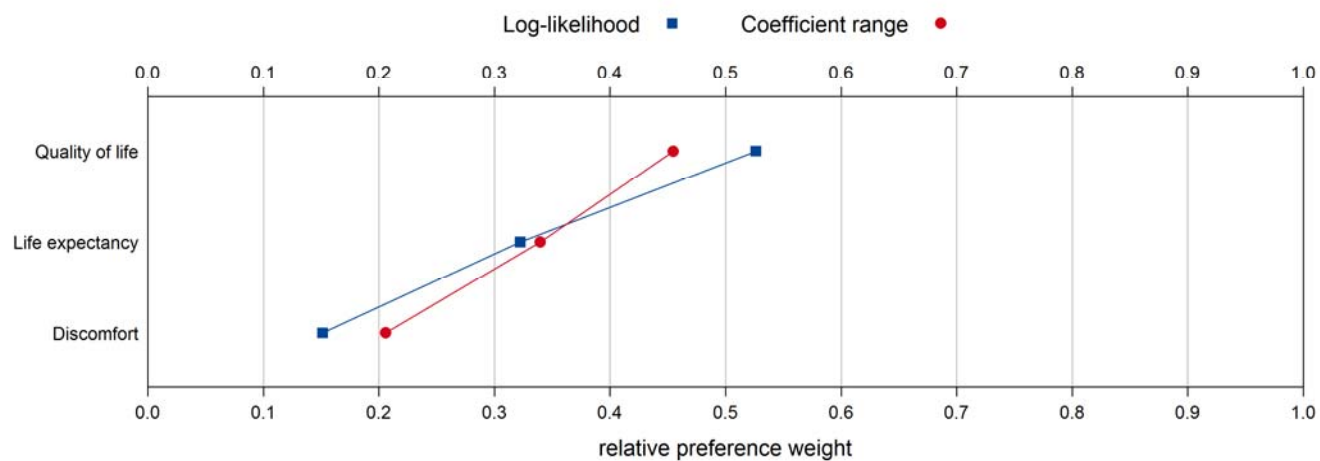
Table 27 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, general population sample (coefficient range method)

Model	Coefficient range	Proportion
Quality of life	0.56	0.42
Discomfort	0.48	0.36
Life expectancy	0.28	0.21

Table 28 – Derivation of the weights for a priori selected criteria in the Therapeutic need domain, decision maker sample (coefficient range method)

Model	Coefficient range	Proportion
Quality of life	0.84	0.45
Life expectancy	0.63	0.34
Discomfort	0.38	0.21

Both methods for calculating the relative preference weights result in similar weights for the Therapeutic need domain (Figure 47 and Figure 48).

**Figure 47 – Comparison of relative preference weights by method in the Therapeutic need domain, general population sample****Figure 48 – Comparison of relative preference weights by method in the Therapeutic need domain, decision maker sample**



5.4.3 Attribute weights in Societal need domain

5.4.3.1 Predictive value of the model

Table 29 shows that the estimated model of the general population and decision maker sample predicts fairly well the observed percentage of choices for each alternative.

Table 29 – Actual and predicted percentage of choice for each alternative

	N	Alternative 1		Alternative 2	
		Actual	Predicted	Actual	Predicted
General population	4288	50.7%	51.4%	49.3%	48.6%
Decision makers	160	47.5%	49.2%	52.5%	50.8%

Table 29 and Table 30 suggest a less than perfect fit of the models to the data. For both samples, the model correctly predicts about 71% of the responses for Societal need.

Table 30 – Societal need: goodness of fit statistics

% of responses correctly predicted by model	
General population	70.7%
Decision makers	72.5%

5.4.3.2 General population model

The summary of the full model results for the general population sample are shown in Table 31.

Table 31 – Societal need: model summary for the general population sample

Attribute	Level	Estimated coefficient ^o	Standard Error	t-value	P-value	Significance level
Prevalence	rare	-0.683	0.043			
	not so frequent	-0.216	0.038	-5.660	<0.001	***
	rather frequent	0.329	0.037	8.793	<0.001	***
	very frequent	0.570	0.039	14.528	<0.001	***
Public expenditure	little public expenditures per patient	-0.521	0.024			
	much public expenditures per patient	0.521	0.019	27.448	<0.001	***

^o Results of a multinomial logistic regression model

*** significant on the 0.1% significance level



The coefficients for rare disease and for not so frequent disease are negative, while those for rather frequent and very frequent disease are positive, meaning that a higher prevalence contributes to a higher perceived societal need. For public expenditures associated with a disease, a higher public expenditure per patient is considered to contribute positively to the societal need.

Table 32 presents the societal need value for all possible scenarios that could be described with the attributes and their levels included in our survey. The higher the value, the higher the societal need is considered by the population. A very frequent disease that induces much additional public expenditures per patient is considered to induce the highest need for developing a better treatment. A better treatment would in this case be a treatment that either reduced public expenditures per patient and/or reduces the prevalence of the disease.

Table 32 – Some examples of conditions with their level of societal need according to the general public

Prevalence	Public expenditure	Societal need score
very frequent	much additional public expenditure	1.090
rather frequent	much additional public expenditure	0.850
not so frequent	much additional public expenditure	0.305
very frequent	little additional public expenditure	0.049
rare	much additional public expenditure	-0.162

For each scenario included in the Societal needs domain, we calculated the probability that an option would be chosen as the highest societal need, out of the full set of possible scenarios. Figure 49 presents the probability that a scenario is chosen as having a higher societal need, out of the full set of all possible scenarios. The probability that a scenario is chosen as the highest societal need increases as the prevalence increases.

Figure 49 – Probabilities of choosing a scenario as having a higher societal need out of the full set of scenarios, general population

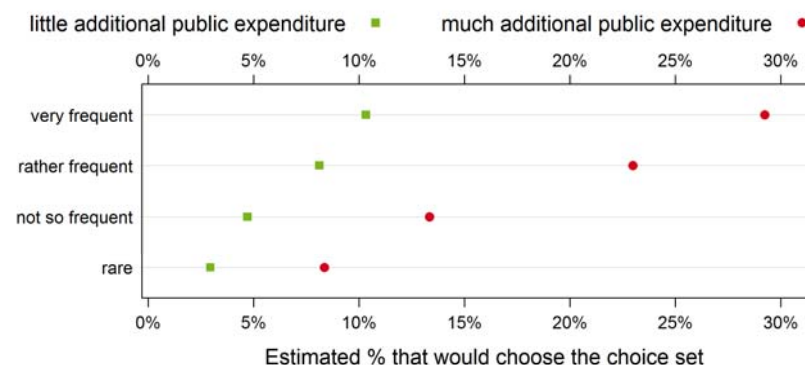


Table 33 shows that societal need becomes much more important when a rare disease becomes more frequent (increase in therapeutic need > 1) than when an already rather frequent disease becomes very frequent (increase in therapeutic need = 0.24).

Table 33 – Differences in coefficients between attribute levels

Prevalence	Rare	Not so frequent	Rather frequent	Very frequent
Rare	0			
Not so frequent	0.47	0		
Rather frequent	1.01	0.55	0	
Very frequent	1.25	0.79	0.24	0



Public expenditure	Little public expenditure	Much public expenditure
Little public expenditure	0	
Much public expenditure	1.04	0

5.4.3.3 Decision makers model

The results for the decision maker sample are presented in Table 34. In contrast to the model for the general public, very few coefficients for “prevalence” are significant in this group. In contrast, the coefficient for public expenditures is highly significant.

Table 34 – Societal need: model summary for the decision maker sample

Attribute	Level	Estimated coefficient ^o	Standard Error	t-value	P-value	Significance level
Prevalence	rare	-0.918	0.223			
	not so frequent	0.126	0.184	0.687	0.492	
	rather frequent	0.222	0.187	1.190	0.234	
	very frequent	0.570	0.196	2.909	0.004	**
Public expenditure	little public expenditures per patient	-0.380	0.113			
	much public expenditures per patient	0.380	0.093	4.087	<0.001	***

^o Results of a multinomial logistic regression model

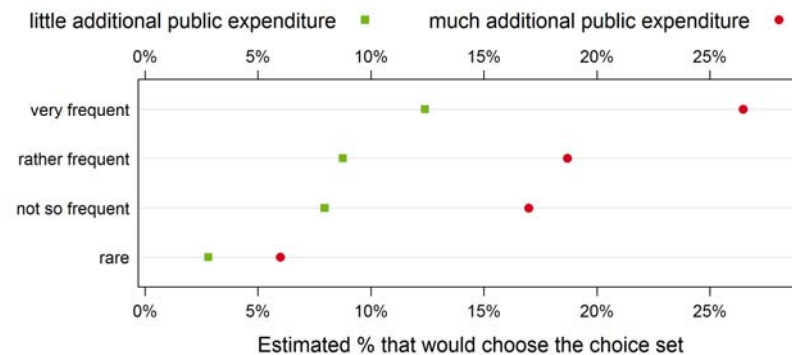
** significant on the 1% significance level

*** significant on the 0.1% significance level



As in the general public model, the predicted probabilities that a particular scenario would be chosen out of the full set of scenarios as having a higher societal need increase as the public expenditures increase, but not as linear as the general public model (Figure 50). The “not so frequent” diseases have a higher probability of being chosen by the decision makers than by the general public.

Figure 50 – Probabilities of choosing a scenario as having a higher societal need out of the full set of scenarios, decision makers



5.4.3.4 Weights for the societal need criteria

The calculations made for estimating the attribute weights according to the log-likelihood approach are shown in Table 35 for the general public and in Table 36 for the decision makers.

Table 35 – Log-likelihood of models in the Societal need domain, general population sample

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes	Likelihood ratio test
Full model	-2329			
Public expenditure excluded	-2778	-448.1	0.648	$X^2[\text{df}=3] = 896.3$ ($p < 0.01$)
Prevalence excluded	-2573	-244.0	0.353	$X^2[\text{df}=1] = 487.9$ ($p < 0.01$)

**Table 36 – Log-likelihood of models in the Societal need domain, decision makers sample**

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes	Likelihood ratio test
Full model	-90.4			
Public expenditure excluded	-102.3	-11.9	0.435	$X^2[df=1] = 23.8$ ($p < 0.01$)
Prevalence excluded	-99.6	-9.1	0.565	$X^2[df=3] = 18.3$ ($p < 0.01$)

The coefficient ranges, used for the calculation of the relative weights of the different attributes by the coefficient range method, are shown in Table 46 for the general population and in Table 47 for the decision makers.

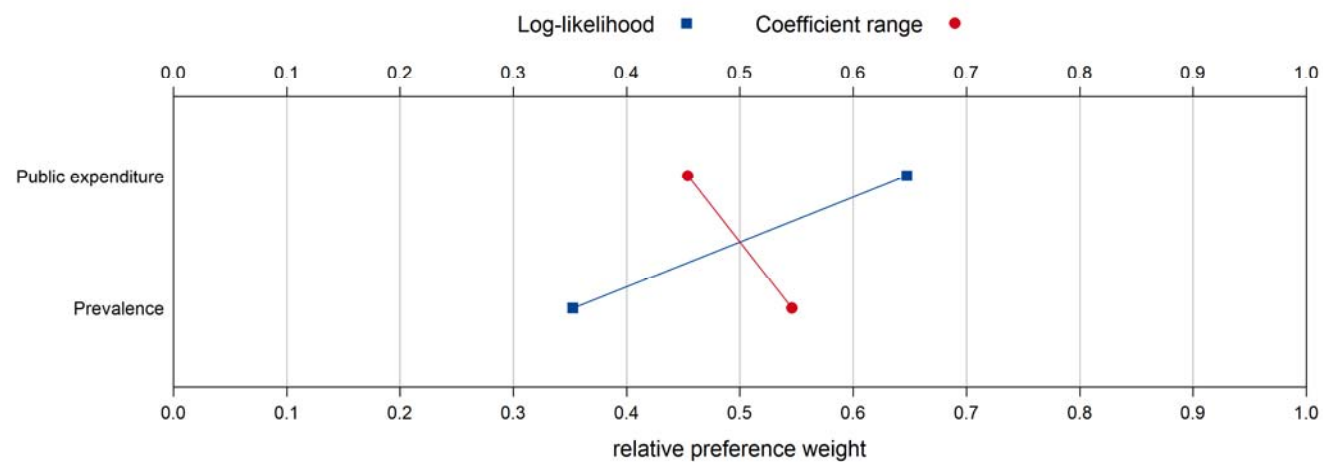
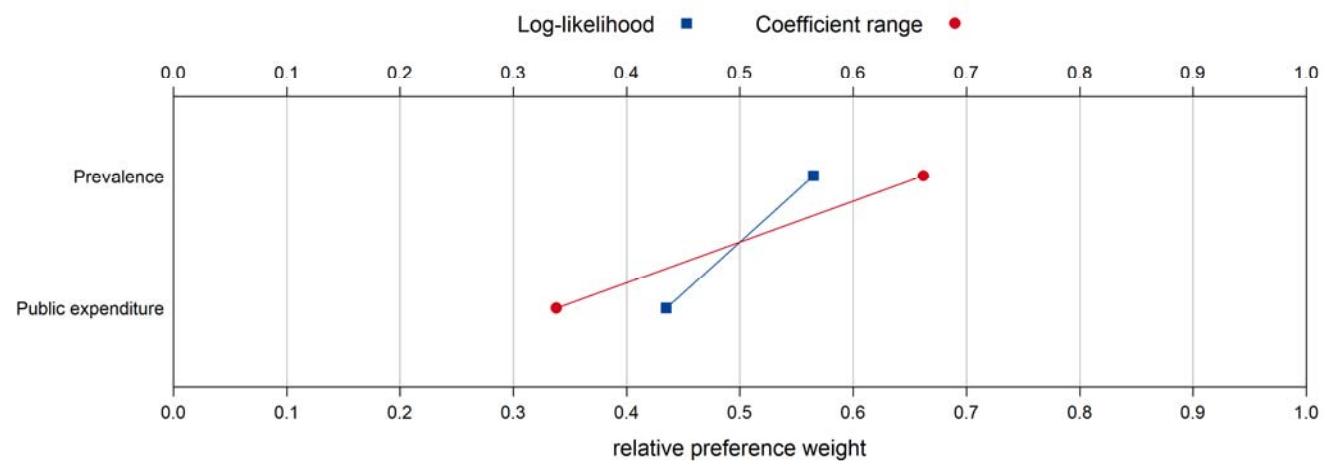
Table 37 – Derivation of the weights for a priori selected criteria in the Societal need domain, general population sample (coefficient range method)

Model	Coefficient range	Proportion
Prevalence	1.25	0.55
Public expenditure	1.04	0.45

Table 38 – Derivation of the weights for a priori selected criteria in the Societal need domain, decision maker sample (coefficient range method)

Model	Coefficient range	Proportion
Prevalence	1.49	0.66
Public expenditure	0.76	0.34

The two methods for calculating the relative preference weights result in different weights for the criteria in the Societal need domain (Figure 51 and Figure 52). In the general population sample, the attributes switch place as to which is the most important attribute. This does not happen in the decision maker sample.

**Figure 51 – Comparison of relative preference weights by method in the Societal need domain, general population sample****Figure 52 – Comparison of relative preference weights by method in the Societal need domain, decision maker sample**



5.4.4 Attribute weights in Added value domain

5.4.4.1 Predictive value of the model

The estimated model of the general population and decision maker sample predicts fairly well the percentage of choices for each alternative (see Table 39).

Table 39 – Actual and predicted percentage of choice for each alternative

	N	Alternative 1		Alternative 2	
		Actual	Predicted	Actual	Predicted
General population	4288	59.5%	60.7%	40.5%	39.3%
Decision makers	160	63.1%	62.9%	36.9%	37.1%

Table 39 and Table 40 suggest a less than perfect fit of the models to the data. For both samples, the model correctly predicts about 80% of the responses for Added value.

Table 40 – Societal need: goodness of fit statistics

	% of responses correctly predicted by model
General population	79.9%
Decision makers	82.2%



5.4.4.2 General population model

The summary of the full model results for the general population sample are shown in Table 41.

Table 41 – Added value: model summary for the general population sample

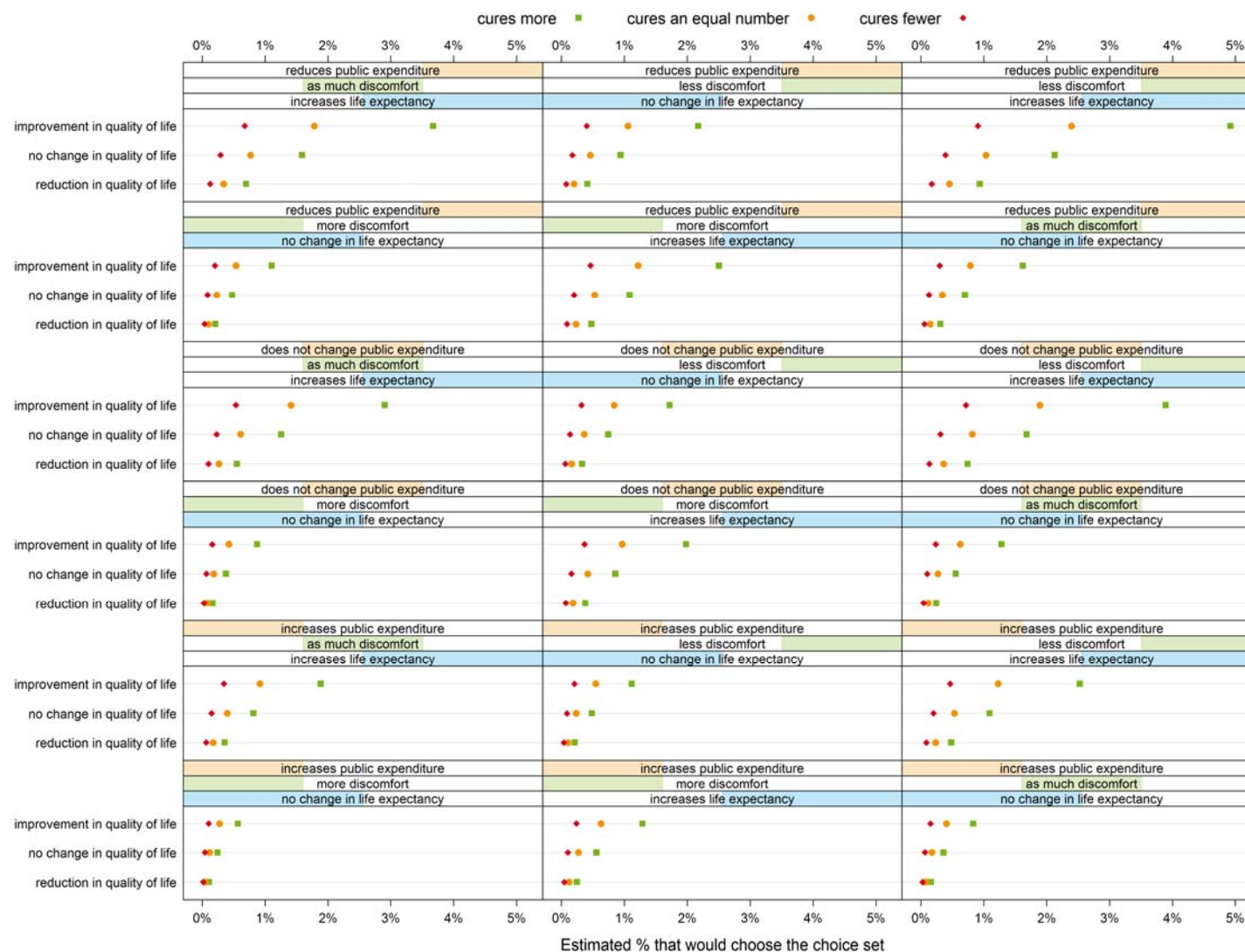
Attribute	Level	Estimated coefficient ^o	Standard Error	t-value	P-value	Significance level
Impact on public expenditure	increases public expenditure	-0.366	0.020			
	does not change public expenditure	0.066	0.018	3.641	<0.001	***
	reduces public expenditure	0.300	0.022	13.897	<0.001	***
Change in quality of life	reduction	-0.826	0.024			
	no change	-0.006	0.018	-0.363	0.717	
	improvement	0.832	0.021	39.129	<0.001	***
Change in life expectancy	does not change	-0.409	0.013			
	increase	0.409	0.013	31.205	<0.001	***
Treatment discomfort	more	-0.353	0.018			
	as much	0.030	0.019	1.611	0.107	
	less	0.323	0.018	18.192	<0.001	***
Change in prevalence	cures fewer	-0.886	0.026			
	cures an equal number	0.082	0.018	4.667	<0.001	***
	cures more	0.804	0.021	38.350	<0.001	***

^o Results of a multinomial logistic regression model

** significant on the 1% significance level

*** significant on the 0.1% significance level

The results show that a reduction in public expenditures, improvement in quality of life, an increase in life expectancy, a reduction in treatment discomfort and a reduction in the number of people with the disease contributes positively to the judgment of the added value of a treatment.





From Table 42 we conclude that the negative effect on the added value of increasing public expenditures is higher (-0.43) than the positive impact of decreasing public expenditures (+0.23). This means that people's preference *against* interventions that increase public expenditures is stronger than their preference *for* interventions that decrease public expenditures. Otherwise stated: **there is less to be gained in terms of added value from choosing a cost-saving intervention than from avoiding a cost increasing intervention**, according to the general public's point of view.

For most other attributes, except for quality of life, a similar observation can be made. For treatment discomfort, **the perceived gain in added value from an intervention that reduces discomfort is lower (+0.29) than the perceived loss in added value from an intervention that increases discomfort (-0.38)**.

Similarly, the **loss in added value from curing less patients (-0.97) is higher than the gain from curing more patients (+0.72)**. Taken together, these observations show that the utility loss associated with something negative (higher expenditures, higher treatment discomfort, less patients cured) is higher than the utility gain associated with something positive (lower expenditures, lower treatment discomfort, more patients cured).

For quality of life, this is less clear: the gain in added value associated with increasing quality of life is about the same as the loss associated with reducing quality of life, disregarding the current quality of life of patients. Not changing quality of life compared to the current situation is considered rather important when the alternative is to reduce quality of life (increase in added value = 0.82).

When comparing the impact of changes on different attributes, it can be observed that improving quality of life (+0.84) is about 2.5 times more important than reducing public expenditures (+0.23) and almost equally important than increasing life expectancy (+0.82). Not changing quality of life compared to the current situation is considered almost twice as important for the added value of an intervention (+0.82) as not changing public expenditures when the alternative is to increase public expenditures (+0.43).

A quality of life reduction can be compensated by a life expectancy increase, all else equal. Thus, an intervention that reduces quality of life but increases life expectancy is valued about the same as an intervention that does not

change quality of life or life expectancy and for which all else is equal (same impact in treatment discomfort, prevalence and public expenditure). It means that people are willing to sacrifice quality of life for a longer life. However, in practice this should be considered in the light of the current quality of life of patients. In our DCE, the current health state was not included in the added value questions. It was therefore impossible to assess the trade-offs made at different levels of baseline quality of life.

Compared to improving quality of life and increasing life expectancy, reducing treatment discomfort is considered relatively less important for the added value of a new treatment according to the general public.

Besides improving quality of life and increasing life expectancy, curing more patients is also considered important for the added value judgment. An intervention that cures more patients has an added value that is 0.72 higher than an intervention that does not change the number of patients cured. This is more than twice as high as the impact of reducing treatment discomfort or reducing public expenditures.

Table 42 – Differences in coefficients between attribute levels

Impact on public expenditures	Increases public expenditures	Does not change public expenditures	Reduces public expenditures
Increases public expenditures	0	-0.43	-0.67
Does not change public expenditures	0.43*	0	-0.23
Reduces public expenditures	0.67	0.23	0

** This figure reflects the difference between the added value of an intervention that does not change public expenditures and the added value of an intervention that does increase the public expenditures. The figure is positive, meaning that the added value of an intervention that does not change public expenditures is higher than the added value of an intervention that increases public expenditures. All figures should be interpreted "ceteris paribus", i.e. all other criteria (quality of life, life expectancy etc) are the same for both interventions.*



Impact on QoL	Reduction	No change	Increase
Reduction	0	-0.82	-1.66
No change	0.82	0	-0.84
Increase	1.66	0.84	0

Impact on life expectancy	No change	Increase
No change	0	-0.82
Increase	0.82	0

Impact on treatment discomfort	More	No change	Less
More	0	-0.38	-0.68
No change	0.38	0	-0.29
Less	0.68	0.29	0

Impact on prevalence	Cures fewer patients	No change	Cures more patients
Cures fewer patients	0	-0.97	-1.69
No change	0.97	0	-0.72
Cures more patients	1.69	0.72	0



5.4.4.3 Decision makers model

The results for the decision maker sample are shown in Table 43.

Table 43 – Added value: model summary for the decision maker sample

Attribute	Level	Estimated coefficient [°]	Standard Error	t-value	P-value	Significance level
Impact on public expenditure	increases public expenditure	-0.499	0.119			
	does not change public expenditure	0.117	0.100	1.168	0.243	
	reduces public expenditure	0.383	0.122	3.141	0.002	**
Change in quality of life	reduction	-1.022	0.150			
	no change	-0.111	0.099	-1.122	0.262	
	improvement	1.133	0.128	8.827	<0.001	***
Change in life expectancy	does not change	-0.643	0.094			
	increase	0.643	0.081	7.925	<0.001	***
Treatment discomfort	more	-0.286	0.087			
	as much	0.079	0.108	0.725	0.468	
	less	0.208	0.100	2.086	0.037	*
Change in prevalence	cures fewer	-0.917	0.154			
	cures an equal number	-0.072	0.101	-0.711	0.477	
	cures more	0.989	0.123	8.066	<0.001	***

[°] Results of a multinomial logistic regression model

* significant on the 5% significance level

** significant on the 1% significance level

*** significant on the 0.1% significance level



5.4.4.4 Weights for the added value criteria

The calculations done for determining the weights for the added value criteria by the log-likelihood method are presented in Table 44 for the general population sample and in Table 45 for the decision maker sample.

Table 44 – Log-likelihood of models in the Added Value domain, general population sample

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes	Likelihood ratio test
Full model	-7684			
Model without change in quality of life	-9124	-1440.0	0.37	$X^2[df=7] = 2880$ ($p < 0.01$)
Model without change in prevalence	-9090	-1406.1	0.36	$X^2[df=7] = 2812.2$ ($p < 0.01$)
Model without change in life expectancy	-8206	-521.9	0.14	$X^2[df=8] = 1043.8$ ($p < 0.01$)
Model without treatment discomfort	-7967	-282.9	0.07	$X^2[df=7] = 565.9$ ($p < 0.01$)
Model without impact on public expenditure	-7927	-242.7	0.06	$X^2[df=7] = 485.4$ ($p < 0.01$)

Table 45 – Log-likelihood of models in the Added value domain, decision makers sample

Model	Log-likelihood	Log-likelihood change	Proportion of log-likelihood change in sum of changes	Likelihood ratio test
Full model	-245.1			
Model without change in quality of life	-316.8	-71.7	0.39	$X^2[df=7] = 143.3$ ($p < 0.01$)
Model without change in prevalence	-298.1	-53.0	0.29	$X^2[df=7] = 106.1$ ($p < 0.01$)
Model without change in life expectancy	-283.0	-37.9	0.21	$X^2[df=8] = 75.7$ ($p < 0.01$)
Model without impact on public expenditure	-259.0	-13.9	0.08	$X^2[df=7] = 27.9$ ($p < 0.01$)
Model without treatment discomfort	-250.4	-5.4	0.03	$X^2[df=7] = 10.7$ ($p < 0.01$)



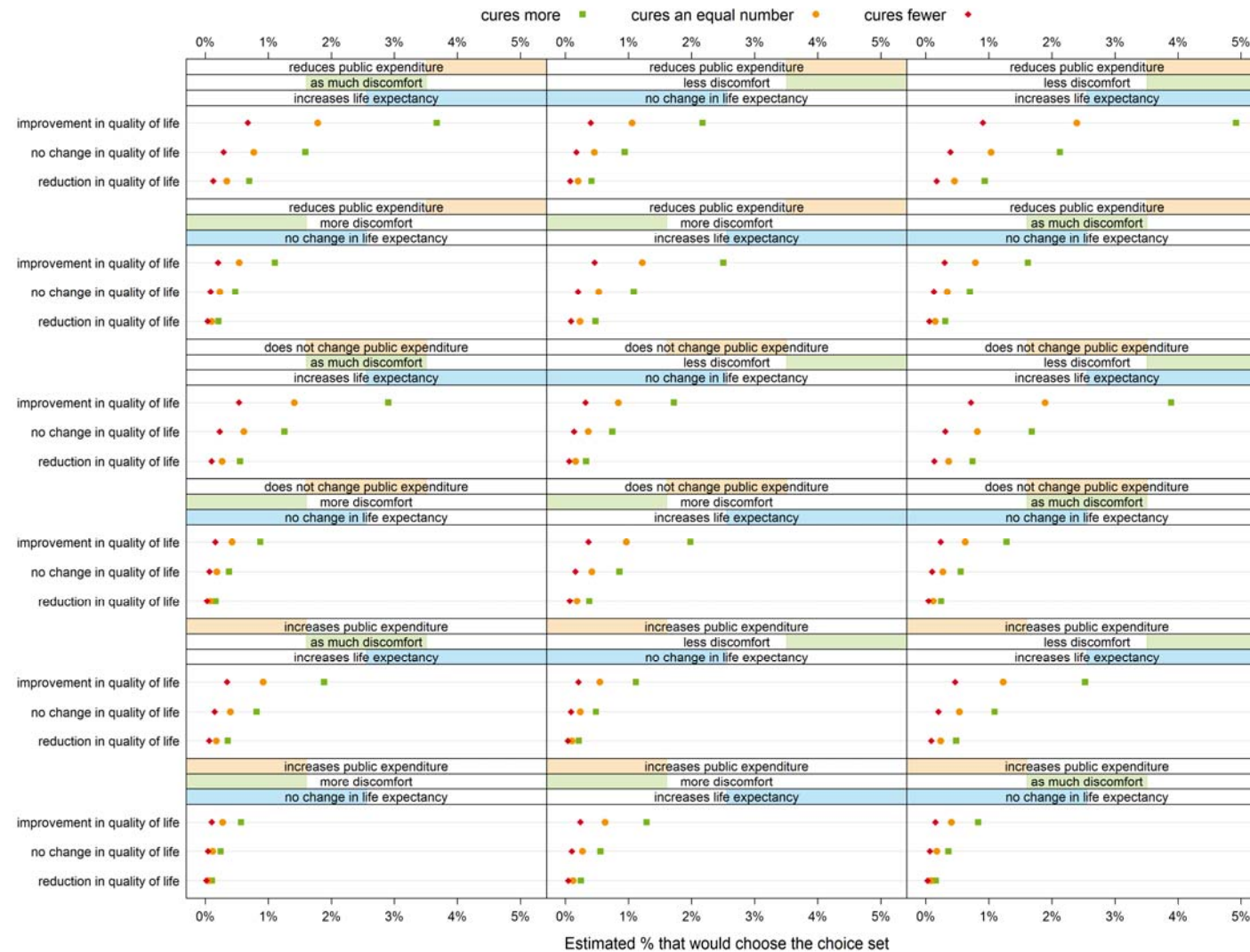
A new treatment's impact on quality of life has the highest importance for both groups, followed by its impact on disease prevalence and impact on life expectancy. For decision makers, a change in prevalence clearly gets a lower weight than a change in quality of life, while for the general population, these are almost of equal importance. Reduction in discomfort gets a very low weight in the decision makers' sample.

Note that the weights for the *improvement* of each of the criteria included to measure therapeutic and societal need are different from the weights of the same criteria in determining the therapeutic and societal need. This is not contradictory, as the added value is assessed independently of the therapeutic need and independently of the societal need, hence independently of disease characteristics. Of course, disease characteristics are important and should be taken into account in the decision-making process (see Chapter 6 on "How to use the results of this study").

Figure 54 presents the probabilities that particular treatment options are considered to have a higher added value out of the full set of treatment options.



Figure 54 – Probabilities of choosing a scenario as having a higher added value out of the full set of scenarios, decision makers





The coefficient ranges, used for the calculation of the relative weights of the different attributes by the coefficient range method, are shown Table 46 in for the general population and in Table 47 for the decision makers.

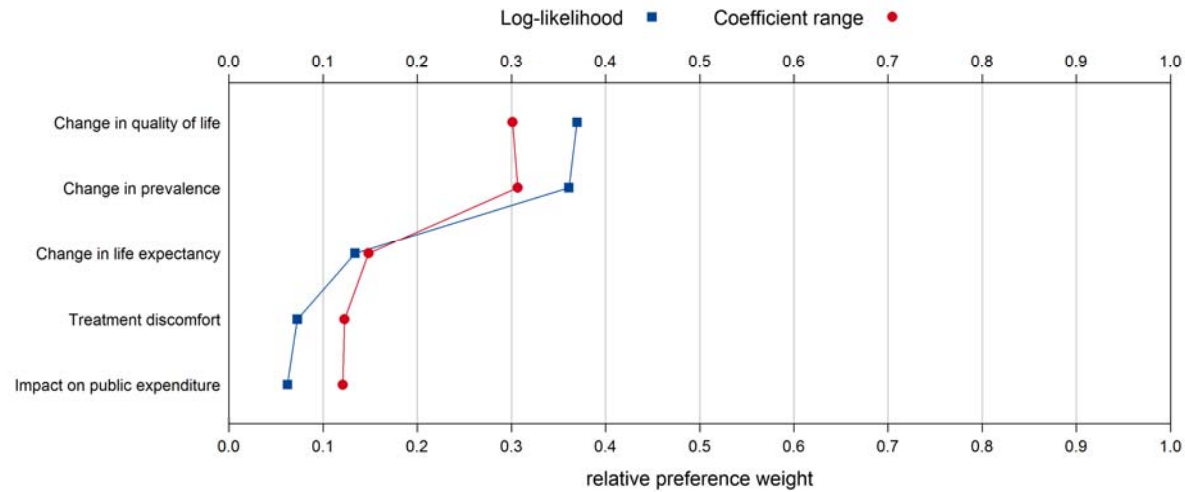
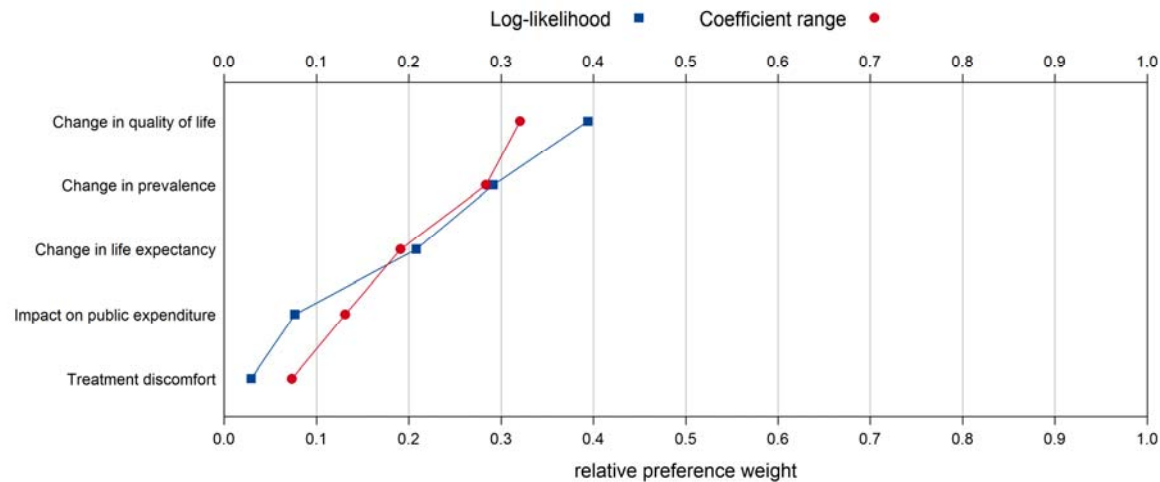
Table 46 – Derivation of the weights for a priori selected criteria in the Added value domain, general population sample (coefficient range method)

Model	Coefficient range	Proportion
Change in prevalence	1.69	0.31
Change in quality of life	1.66	0.30
Change in life expectancy	0.82	0.15
Impact on treatment discomfort	0.68	0.12
Impact on public expenditure	0.67	0.12

Table 47 – Derivation of the weights for a priori selected criteria in the Added value domain, decision maker sample (coefficient range method)

Model	Coefficient range	Proportion
Change in quality of life	2.16	0.32
Change in prevalence	1.91	0.28
Change in life expectancy	1.29	0.19
Impact on public expenditure	0.88	0.13
Impact on treatment discomfort	0.49	0.07

The coefficient ranges method and the log-likelihood method result in a similar ranking of the weights for the Added value domain (Figure 55 and Figure 56), except for the changes in quality of life and changes in prevalence in the general population sample. However, in both methods the weights for both criteria are highly similar.

**Figure 55 – Comparison of relative preference weights by method in the Added value domain, general population sample****Figure 56 – Comparison of relative preference weights by method in the Added value domain, decision maker sample**



Key points

- Within each domain, more than 75% of respondents reported to be certain about their choice. Respondents needing more reminders tend to be more uncertain of their choices.

Therapeutic need

- The predictive value of the estimated model of the general population and decision maker samples is good but not perfect. The model correctly predicts three quarter of the responses for both samples.
- Age is an important but not absolute criterion for judging therapeutic need. In general, older people with a good quality of life and little treatment discomfort are considered to have a lower therapeutic need than younger people with a bad quality of or high treatment discomfort. However, when older patients have a bad quality of life and high discomfort they are considered to have a higher therapeutic need than younger patients with a good quality of life and little treatment discomfort.
- People do not discriminate very clearly between the very young (<18 years of age) and the working age adults (18-64 years of age) when judging therapeutic need
- People place a higher value on avoiding quality of life loss in patients that currently have a rather good quality of life than on avoiding further loss in quality of life in patients who are already in a bad quality of life state.
- Respondents did not clearly discriminate between “die almost immediately” and “die 5 years earlier”; which suggests that they dichotomized this attribute into “lethal” and “non-lethal” disease.

- Respondents trade-off quality of life and life expectancy for the judgment of therapeutic need. A lower quality of life and a low impact on life expectancy are equivalent to a better quality of life and a higher impact on life expectancy
- According to the general public the impact of a disease on quality of life and the impact of the current treatment’s discomfort should weight more in the assessment of therapeutic need than life expectancy.
- In contrast to the general public, decision makers attach more importance to the impact of a disease on life expectancy than to the discomfort of current treatment, when judging therapeutic need.

Societal need

- The predictive value of the estimated model of the general population and decision maker samples is good but not perfect. The model correctly predicts about 71% of the responses for both samples.
- A very frequent disease that induces much additional public expenditures per patient is considered to induce the highest need for developing a better treatment.
- Less frequent diseases have a higher chance of being chosen by the decision makers than by the general public.
- According to the general public the impact of a disease on public expenditures per patient should weight more in the judgment of societal need than the prevalence of the disease.
- According to decision makers, prevalence should weight more than public expenditures per patient.

Added value

- The predictive value of the estimated model of the general population and decision maker samples is good but not perfect. The model correctly predicts about 80% of the responses for both samples.



- **Not surprisingly, an intervention that reduces public expenditures per patient, improves quality of life, increases in life expectancy, reduces treatment discomfort and reduced the prevalence of the disease has the highest added value.**
- **The added value gain of lower public expenditures, a lower discomfort of treatment or a reduction in disease prevalence is lower than the added value loss associated with higher expenditures per patient, higher treatment discomfort or fewer patients cured. In other words, the value loss associated with something negative (e.g. higher expenditures) is higher than the value gain associated with something positive (e.g. reduced expenditures).**
- **Both the general public and the decision makers give the highest weight to the impact of a new intervention on quality of life when judging the added value of that intervention.**
- **Quality of life is followed by impact of the new treatment on the prevalence of the disease and on life expectancy.**
- **The impact on treatment discomfort and public expenditures per patient are considered of lowest importance by both the general public and the decision makers.**

5.5 Comparison of the weights of subsamples of the general population

We performed the same analysis as for the total general population sample on subgroups of the general public, defined by age category, self-reported health status and number of reminders received. The objective of these analyses was to examine whether preferences differed between population subgroups and to assess the risk of bias in our results if our survey sample would not be representative. Unfortunately, it has not been possible for all variables of interest to check whether our survey is representative, because there are no national data available (e.g. for self-reported health status) but only proxies with their own limitations.

If the results of subsamples are similar to those of the complete sample, the risk of bias due to lack of representativeness can be expected to be lower. The results might still be biased, though, because people with other preferences might simply not have participated in our survey. The non-response rate was 76%. There is no direct way to test whether non-responders have different preferences than responders. The indirect way to assess this is to compare the preferences of the early responders with those of the late responders. The (untested) hypothesis is that late responders are more alike non-responders than early responders.

The comparisons are made for the weights as obtained with the log-likelihood approach. Especially differences between sub-groups in the order of criteria are of interest, as these have the biggest impact on the outcome of an MCDA. Statistical significance of differences in point estimates between subgroups are less relevant. Due to small numbers in subgroups, confidence intervals are likely to be large and differences between point estimates statistically not significant. However, for the application of the weights in MCDA, the ranking of the criteria matters more than which numerical value is given to that criterion. The relative values are important, but not the absolute values. The discussion will therefore focus on differences in the order of criteria between sub-groups.



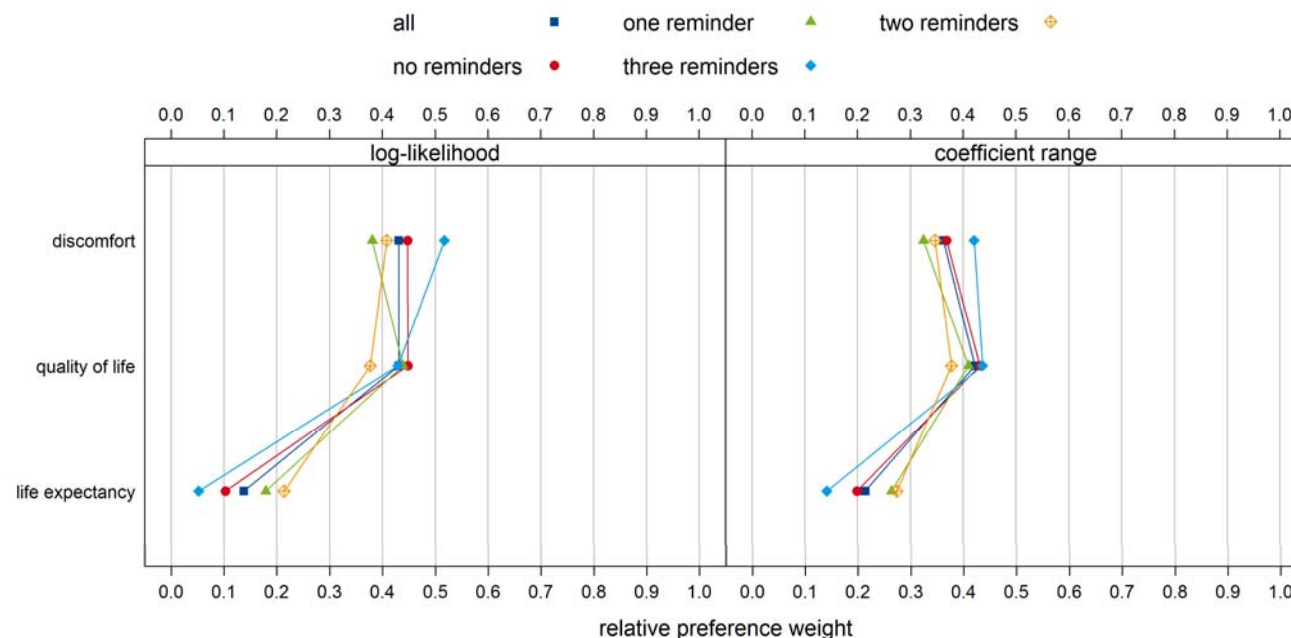
5.5.1 Weights by subgroup defined by number of reminders received

We compared the relative weights of criteria between the groups of respondents who responded immediately after the first invitation or after the

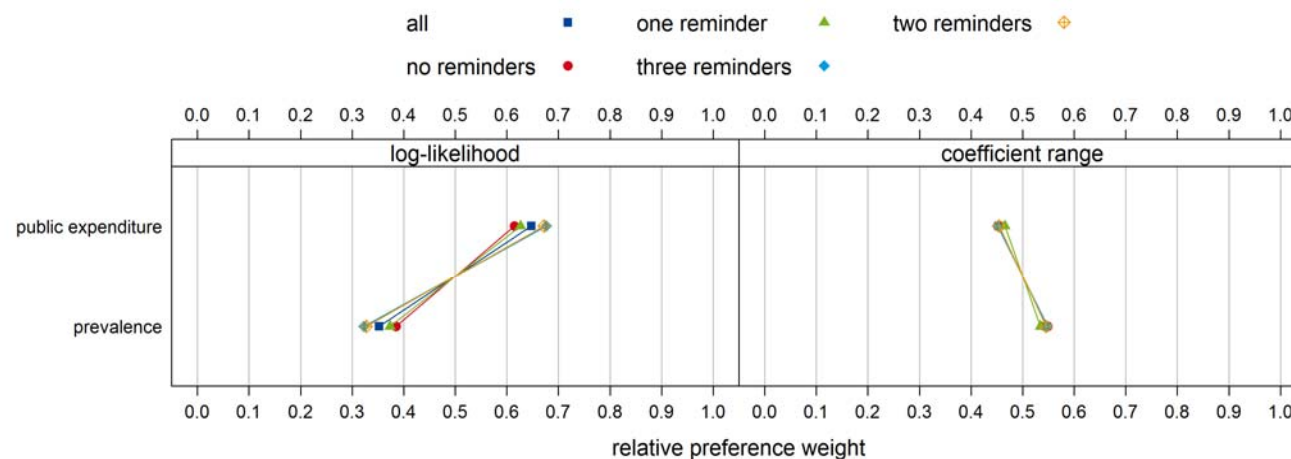
first, second or third reminder. The weights of the therapeutic need criteria for the different subgroups are presented in Figure 57.

There is no difference between the weights for therapeutic need according to the number of reminders received, and especially not between the full sample results and the late responder results.

Figure 57 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of number of reminders received



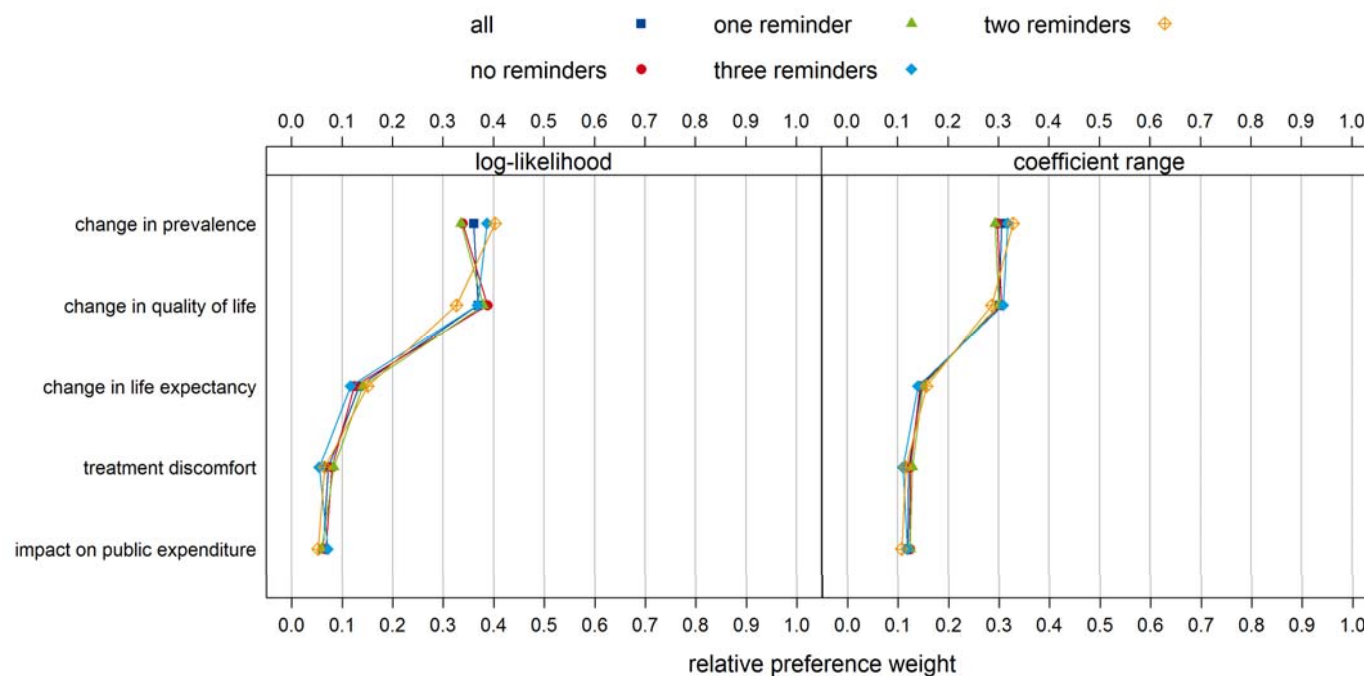
Similarly, for societal need, no differences were observed between the full sample and the late responders (Figure 58). Disease-related public expenditure is considered more important than prevalence by all subgroups..

**Figure 58 – Relative weights of decision criteria for societal need by subgroup defined in function of number of reminders received**

For added value, late responders (two or three reminders) seem to attach slightly more importance to the impact of a treatment on prevalence than to the impact on quality of life, whereas the other groups attach slightly more importance to impact on quality of life than to impact on prevalence (Figure 59). The group that had received two reminders also valued impact on treatment discomfort as less important than impact of the disease on public expenditure, contrary to the other groups.



Figure 59 – Relative weights of decision criteria for added value by subgroup defined in function of number of reminders received

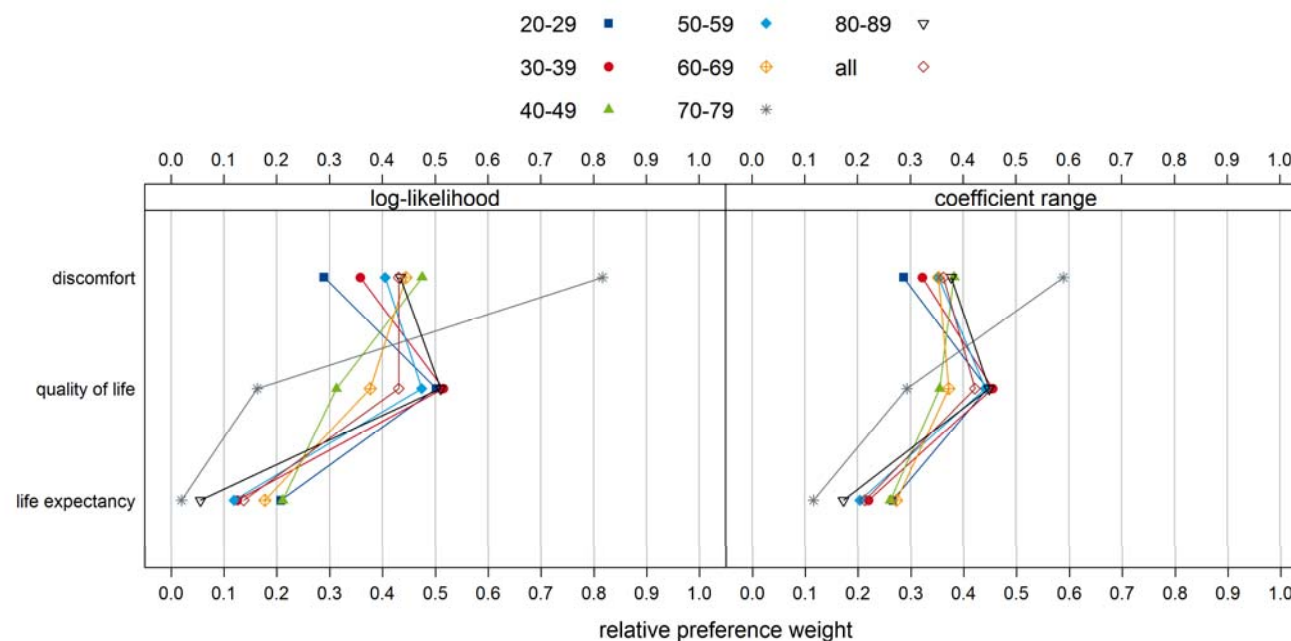


5.5.2 Weights by subgroup defined by age

When subgroups defined by age category of the respondent are compared, we observe that the 80-89 year olds have different preferences than the other age groups. It should be noted, however, that the estimates for this age group are based on 121 participants only, whereas the other age categories had between 331 and 963 participants.

For therapeutic need, respondents between 70 and 79 years of age give much more importance to the criterion of discomfort of current treatment and less to the criterion of quality of life under current treatment than the other age groups.

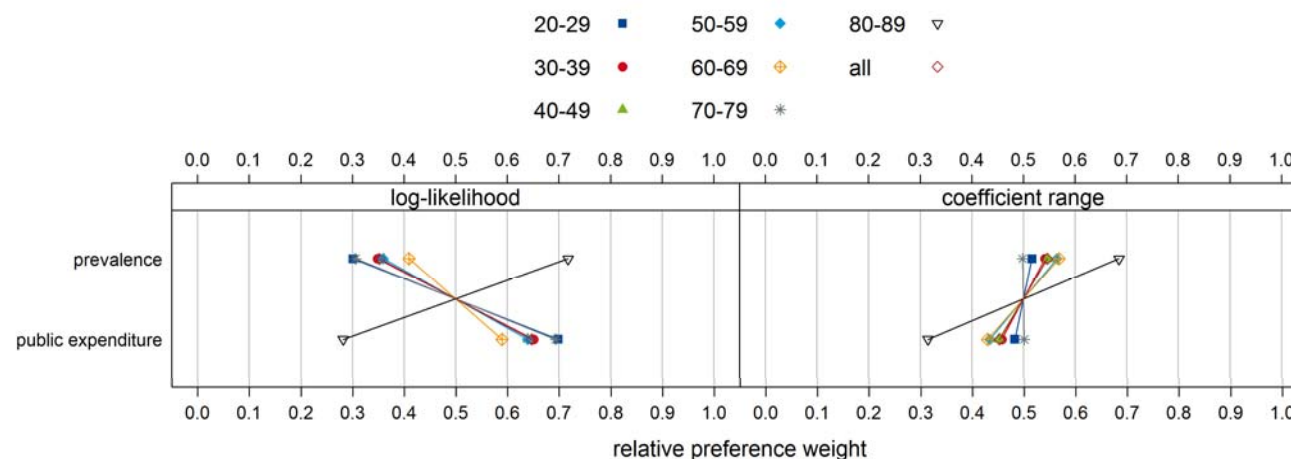
Three age groups consider discomfort of current treatment as more important than quality of life: 40-49 year olds, 60-69 year olds and 80-89 year olds (Figure 60). For all other age groups, quality of life is more important than discomfort of current treatment. Because of these opposing preferences across age groups, the model results in almost equal weights for quality of life and discomfort for the full sample.

**Figure 60 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of age**

For societal need as well we observe opposing preferences for the 80-89 year olds as compared to the other age groups (Figure 61). Whereas all other age groups give a higher weight to public expenditures in judging societal need, the 80-89 year olds give a higher weight to prevalence. The overall estimates seem to reflect more closely the preferences of the 30-59 year olds than of the 20-29 or the 60-89 year olds.

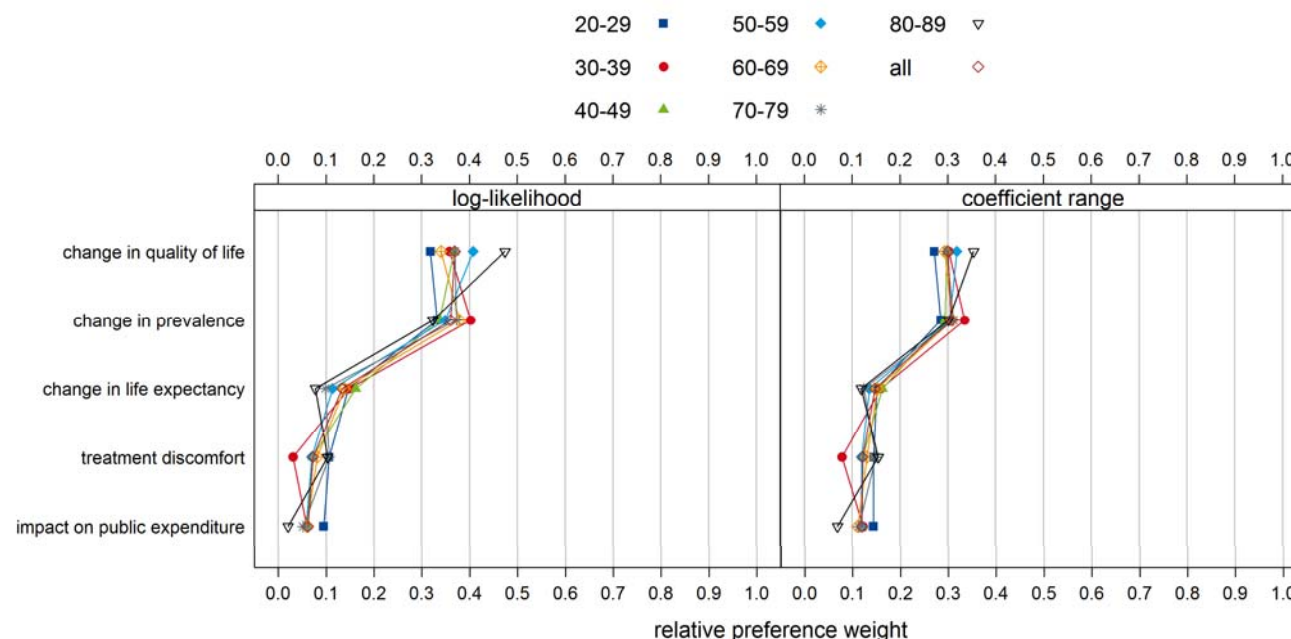


Figure 61 – Relative weights of decision criteria for societal need by subgroup defined in function of age



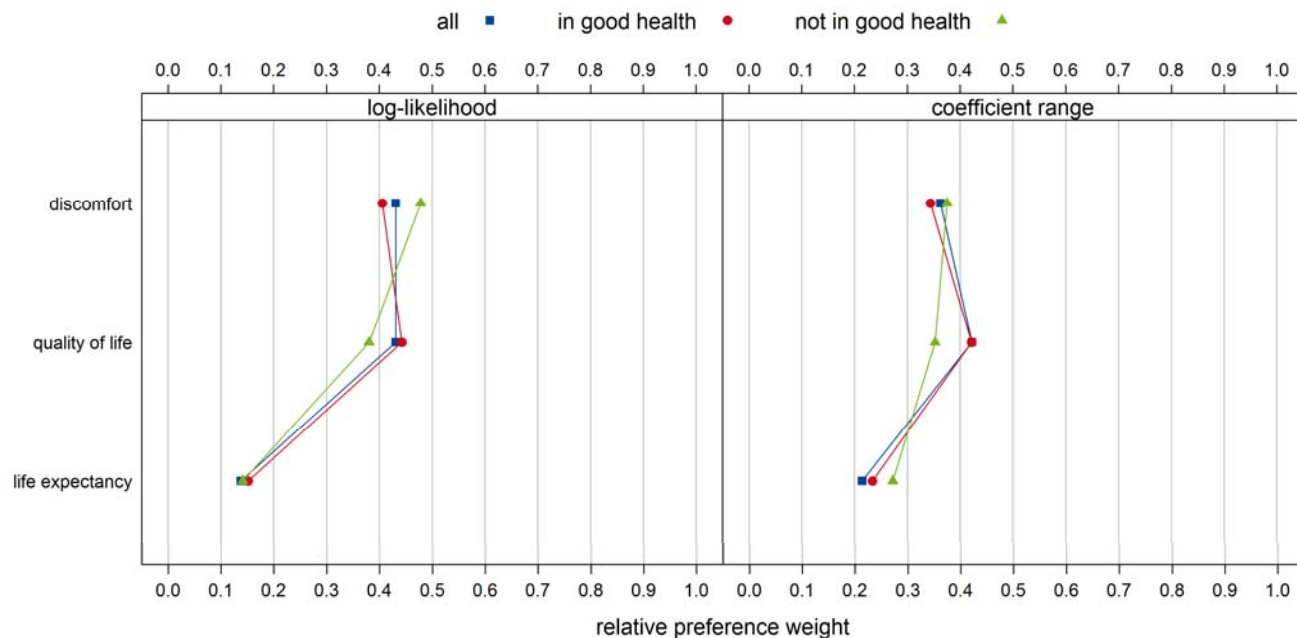
As for the judgment of the added value of new treatments, the elderly (80-89 years of age) seem to give relatively more weight to improvements in quality of life than the other age groups (Figure 62). At the same time, changes in treatment comfort are more important than changes in life expectancy for this group as well as for the 70-79 years old. This means that this age groups values living better more than living longer, whether “better life” is defined by better quality of life or less treatment discomfort. Again, the significance of these differences is questionable, given the low number of respondents in these higher age groups.

In contrast to the older age groups, the other age groups typically give more weight to improvements in life expectancy than to reductions in discomfort, but they also give more weight to improvements in quality of life than to increases in life expectancy. The respondents in the youngest age group (20-29y) give relatively more weight to reductions in public expenditures compared to the other age groups, although this criterion also for this age group remains the least important for the assessment of the added value of a new intervention.

**Figure 62 – Relative weights of decision criteria for added value by subgroup defined in function of age**

5.5.3 Weights by subgroup defined by self-reported health status

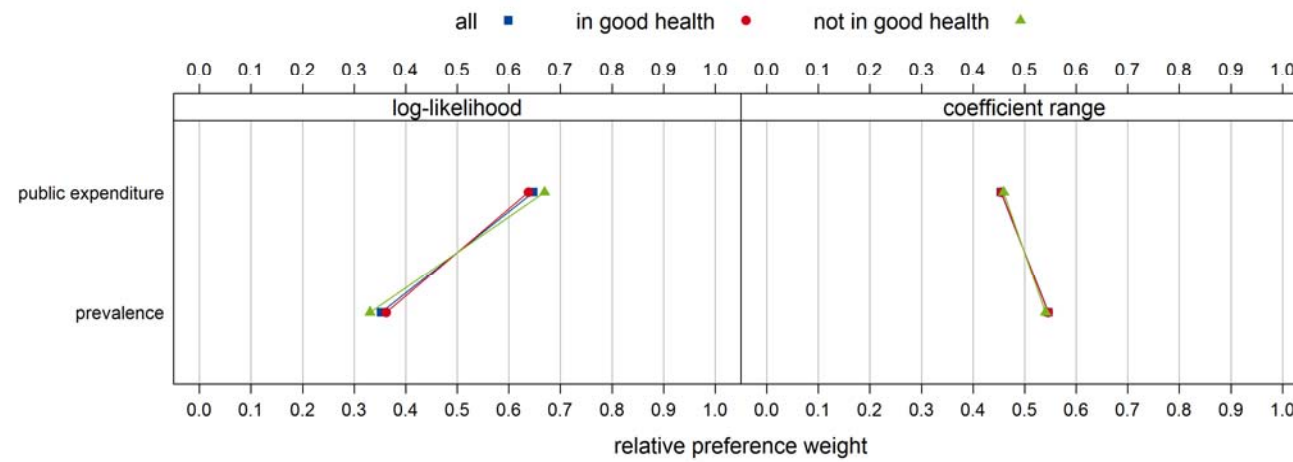
Figure 63 shows the weights of the subgroups of respondents that described their current health as “good” or “not good”. People who report being currently in good health give slightly more weight to quality of life under current treatment when judging therapeutic need than to discomfort of current treatment. This is opposite to the weights given to these criteria by respondents who report not being in good health. These patients find it more important to reduce treatment discomfort than to increase overall quality of life. Both subgroups give the lowest weight to reductions in life expectancy due to the disease.

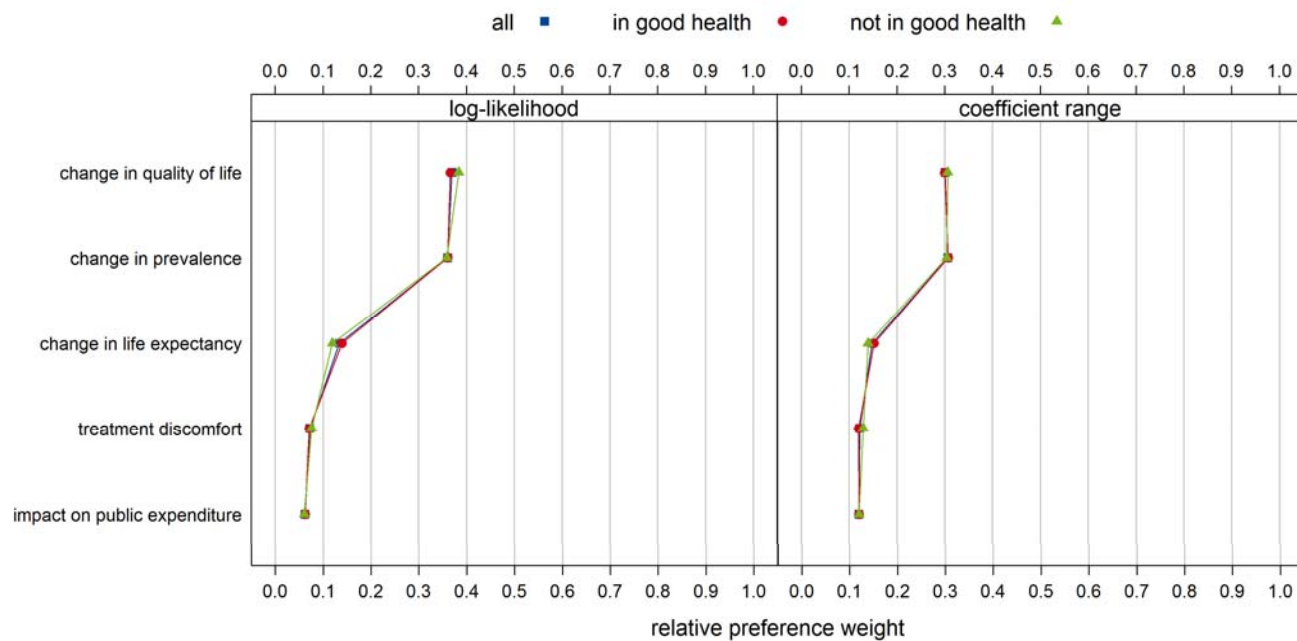
**Figure 63 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of self-reported health status**

For societal need and for added value, there is no difference between the preferences of respondents in good self-reported health and the preferences of respondents in bad self-reported health (see Figure 64 for societal need and Figure 65 for added value).



Figure 64 – Relative weights of decision criteria for societal need by subgroup defined in function of self-reported health status



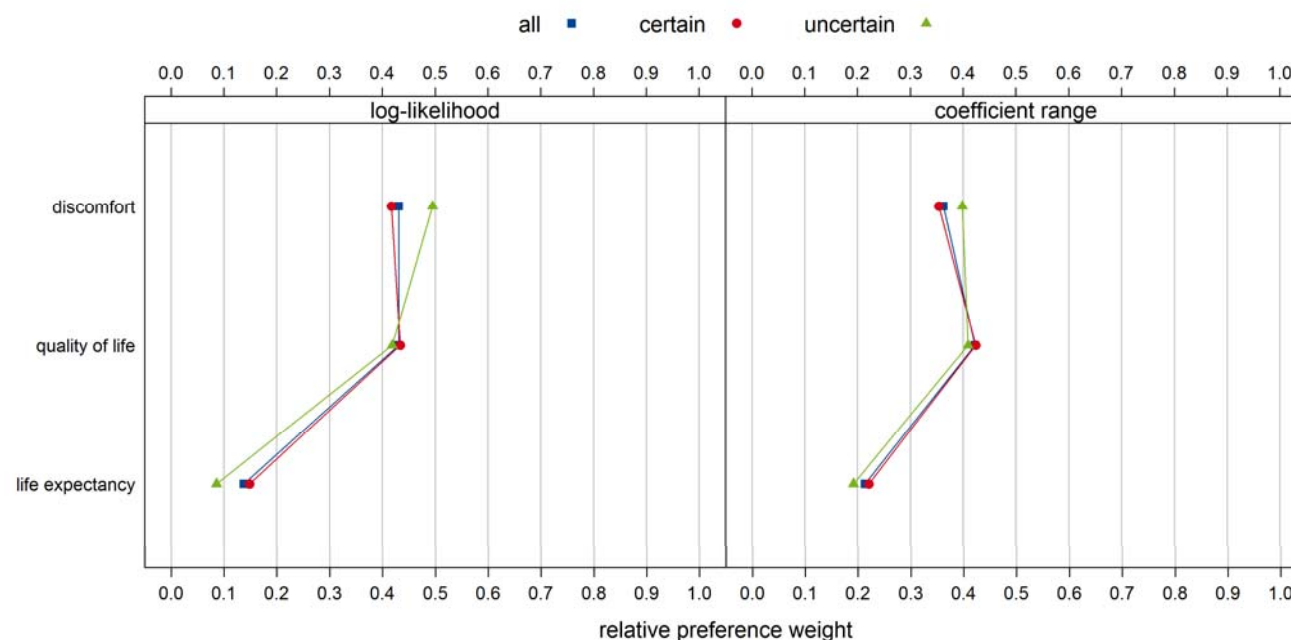
**Figure 65 – Relative weights of decision criteria for added value need by subgroup defined in function of self-reported health status**

5.5.4 Weights by subgroup defined by uncertainty

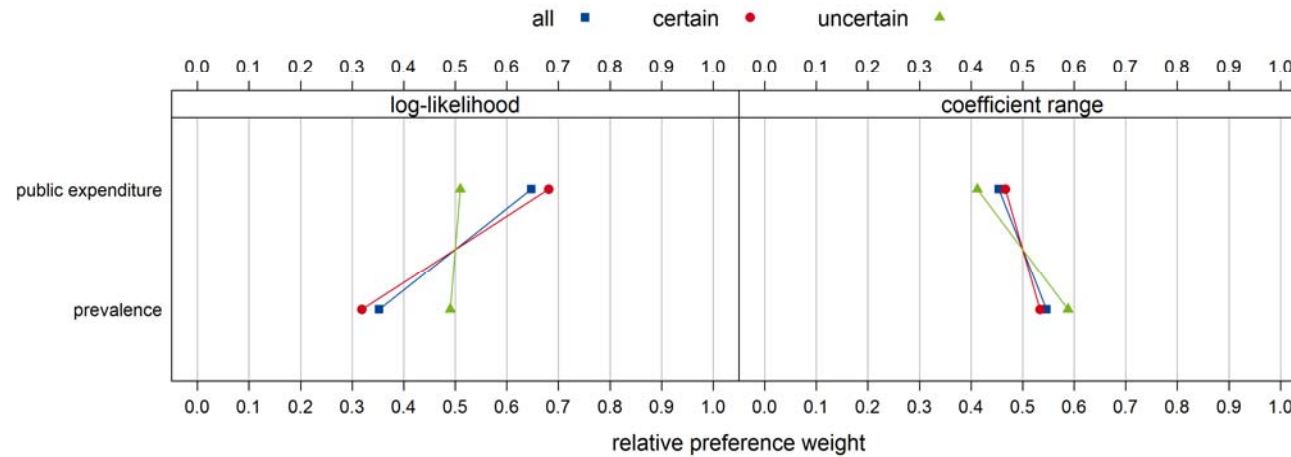
Relative weights of criteria determining therapeutic need do not differ between groups who are uncertain about their responses and groups who are certain about their responses (Figure 63).



Figure 66 – Relative weights of decision criteria for therapeutic need by subgroup defined in function of certainty of responses



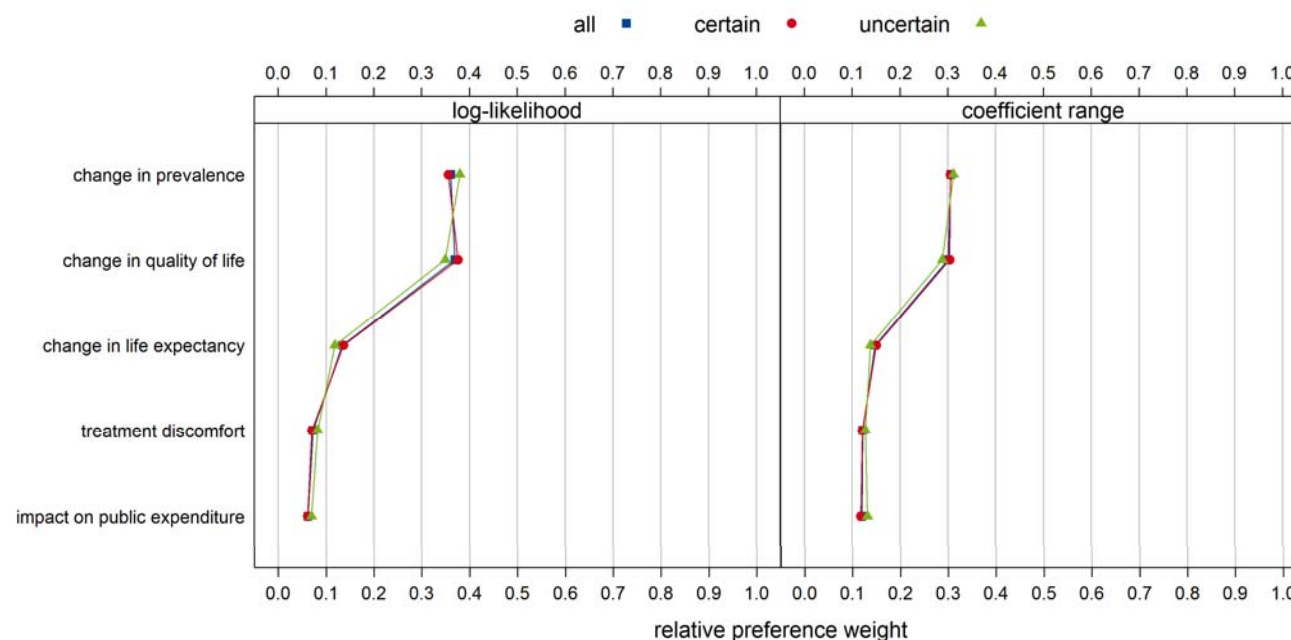
For societal need, however, there is a marked difference between the weights given to the criteria by the group of respondents that is uncertain about its responses and the group of respondents who is certain about its responses (Figure 67). In fact, the uncertainty about which criteria matters most is reflected in the weights of the uncertain respondents. The weight for public expenditures induced by the disease is almost equal to the weight of the prevalence of the disease. Respondents who are certain about their responses clearly gave more weight to public expenditures for judging societal need.

**Figure 67 – Relative weights of decision criteria for societal need by subgroup defined in function of certainty of responses**

Preferences for added value criteria were similar between the groups, although people who were uncertain about their responses gave slightly more weight to changes in prevalence than to changes in quality of life, contrary to than people who were certain about their responses.



Figure 68 – Relative weights of decision criteria for added value by subgroup defined in function of certainty of responses



Key points

- There is no difference between the weights for therapeutic and societal need or added value between late responders and the full sample.
- No differences in preference weights for therapeutic and societal need or added value were observed between subgroups defined by self-reported health status.
- Preference weights for therapeutic need and societal need differ across age subgroups. The results should be treated cautiously, however, as the results for the group of 70-79 year olds is based on only 121 responses.
- For therapeutic need, respondents between 70 and 79 years of age give more importance to the criterion of discomfort of current treatment and less to the criterion of quality of life under current treatment than the other age groups.
- For societal need, the oldest age group (80-89) gives a higher weight to prevalence compared to the full sample.
- For added value, the preferences of the older age groups are generally in line with those of the full sample, although the order is slightly different: changes in discomfort should weight more than changes in life expectancy according to the 80-89 year olds and changes in quality of life should weight more than changes in prevalence.



6 FUTURE USE THE RESULTS OF THIS STUDY

Although additional steps are needed before the results of this study can start to be used in practice, we will briefly describe the general framework of their possible future use.

6.1 Practical steps in the MCDA

The application of MCDA in real life would require the following steps:

Step 1: Scoring the evidence described in the assessment report for each criterion

The consequences of each option are to be described in a Health Technology Assessment report: no value judgments, only description of evidence, uncertainty of evidence and evidence gaps related to each of the criteria identified in Table 48.

Options are to be scored on the criteria by the advisory commission members using a pre-determined scale. Scores should be based on the evidence provided in the assessment report and – if necessary – input from patients (e.g. for the scoring of quality of life with current treatment and discomfort of current treatment). Guidance on how to translate evidence into a score has yet to be developed. This is planned for 2015.



Table 48 –Scoring table

Criterion	Evidence	Score
Quality of life under current treatment	Summarizes the results of studies about the impact of the disease on quality of life under current treatment.	Committee members score the evidence on a pre-determined scale
Life expectancy under current treatment	Summarizes the results of studies about the impact of the disease on life expectancy under current treatment.	Committee members score the evidence
Discomfort of current treatment	Summarizes the results of studies about the discomfort of current treatment.	Committee members score the evidence
Prevalence	Shows figures on the prevalence of the disease in Belgium.	Committee members score the evidence
Disease-related public expenditure per patient	Summarizes the results of studies about the public expenditure related to the disease differentiated by sector: healthcare related costs, invalidity, workdays lost, etc.	Committee members score the evidence
Improvement in quality of life of new treatment compared to current treatment	Summarizes the results of studies about the incremental effect of the new treatment on quality of life compared to the current treatment.	Committee members score the evidence
Improvement in life expectancy of new treatment compared to current treatment	Summarizes the results of studies about the incremental effect of the new treatment on life expectancy compared to the current treatment.	Committee members score the evidence
Reduction of treatment discomfort	Summarizes the results of studies about the incremental effect of the new treatment on treatment discomfort compared to the current treatment.	Committee members score the evidence
Reduction in disease-related public expenditure per patient	Summarizes the results of studies about the incremental effect of the new treatment on public expenditures per patient compared to the current treatment.	Committee members score the evidence
Reduction in prevalence of treatment	Summarizes the results of studies about the reduction of disease prevalence due to the new treatment compared to the current treatment.	Committee members score the evidence



Step 2: Weighting the scores with the public preference weights

The scores given to each criterion should be weighted with the weights presented in the current study (Table 49). The weights prescribe to what extent a particular score should weigh in the appraisal of therapeutic need, societal need or added value.

Table 49 – Weights of decision criteria per domain as measured in the general public

Domain	Decision criteria	Weights*
Therapeutic need	Discomfort of current treatment	0.43 – 0.36
	Quality of life with current treatment	0.43 – 0.42
	Life expectancy despite current treatment	0.14 – 0.22
Societal need	Disease-related public expenditures per patient	0.65 – 0.45
	Prevalence of the disease	0.35 – 0.55
Added value of new treatment	Impact on quality of life	0.37 – 0.30
	Impact on prevalence of the disease	0.36 – 0.31
	Impact on life expectancy	0.14 – 0.15
	Impact on discomfort of treatment	0.06 – 0.12
	Impact on disease-related public expenditure per patient	0.07 – 0.12

* The first figure is the weight as derived with the log-likelihood method, the second figure is the weight derived with the coefficient range method.

In contrast to the examples in literature, the weights measured in the current study are based on population preferences, instead of preferences derived from commission members or a non-representative sample of the public. To avoid random modifications of the weights, as a consequence of which the consistency between decisions could be reduced, we recommend not to

modify the weights but take other considerations into account after the application of the MCDA (Step 5). If there is any reason to believe the outcome of the MCDA exercise is unacceptable, this should be explained by other considerations and not by wrong weights for the attributes in a particular case.

Step 3: Calculating the weighted sum of the scores per domain

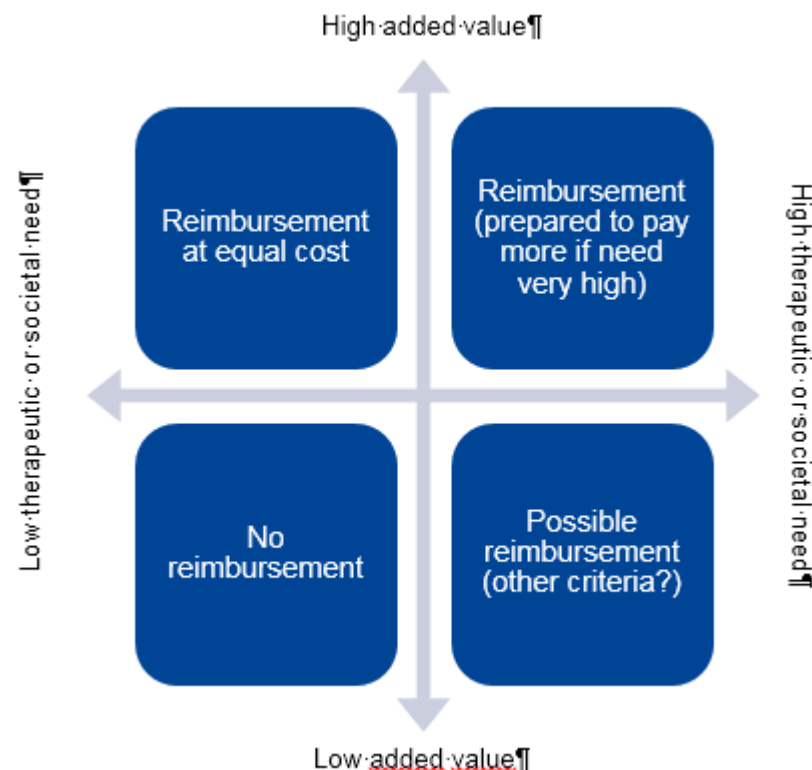
A weighted sum of scores relating to all criteria within one domain should be calculated to obtain the domain score. The MCDA process as presented here will lead to **three** weighted scores per assessed intervention:

- a weighted score for therapeutic need
- a weighted score for societal need
- a weighted score for added value.

The three scores should be compared with the weighted scores of previously assessed diseases (in case of need) or previously assessed health interventions (in case of added value). In other words, the weighted score as such is insufficient to allow a judgment about need or added value; it is only relative to the weighted scores of other diseases or interventions that a judgment will become possible.

Step 4: Interpreting and using the results on the three domains

Once the three sums of weighted scores have been calculated, the commission has to consider in which quadrant of Figure 69 the intervention is located. The higher the total weighted score on therapeutic need, the higher the perceived therapeutic need. The same reasoning applies to all three domains.

**Figure 69 – Preparedness to pay (more) for a new intervention****Step 5: Making a decision**

A final step is making the decision, based on the ranking of the new intervention and its targeted condition on therapeutic need, societal need and added value.

Systematic application of the MCDA framework in decision making will eventually give rise to three MCDA “league tables”: one for therapeutic need, one for societal need and one for added value. The league table is an ordered list of therapeutic and societal needs and of interventions’ added value in function of their level of priority. The higher the score resulting from the MCDA, the higher the therapeutic or societal need or added value of an

intervention and the more likely the society is prepared to pay for the intervention.

The response to the fourth question of the decision-making framework (preparedness to pay more) depends on *both* the place of the intervention in the added value league table and the place of the patient group, respectively disease, in the therapeutic need, respectively societal need, league table. It is more likely that society is prepared to pay more for a new intervention if both the therapeutic and societal need are high and the added value is considerable (Figure 69).

In case of a low therapeutic and societal need and a high added value, the authorities might still want to reimburse a new intervention, but only if the overall cost of the treatment is the same as that of the comparator. An economic evaluation can provide this information.

Interventions with a low added value and a low need do not offer value for money and should therefore not be reimbursed. In cases where the therapeutic or societal need is high and the added value low, decision makers might be more inclined not to reimburse a new intervention. However, in some particular cases, e.g. when no active alternative treatment is available but the only available alternative is best supportive care, they still might decide to reimburse the new intervention to keep the door open for further improvements in the development of the intervention. In such cases, specific conditions for reimbursement are often defined (i.e. who gets reimbursement, under which conditions) and a re-assessment after some time is scheduled. These conditions and additional requirements are needed to ensure that the reason for initially granting reimbursement continue to apply.

In the decision making phase it might be necessary to take other considerations into account that are felt to be relevant but are not yet covered by the criteria. These additional considerations can give rise to modifications in the ranking of a disease (in case of need) or intervention (in case of added value). It should be documented what these additional considerations are and how they modified the ranking on therapeutic need, societal need and/or added value. Additional *criteria* should only be criteria that have not been considered already in the MCDA. If not, the whole point of the MCDA is lost. A more extensive discussion on considerations beyond MCDA is provided in the next paragraph.



6.2 Considerations beyond the MCDA

The decision process does not stop here though. It has been stressed before on several occasions in this report that **the MCDA tool is an aid to decision making**, not a guideline or magic formula. Crucial for any policymaker, making decisions on behalf of and for the benefit of society, are the **ethical considerations**. Policymakers are elected for applying general ethical principles when making decisions. The ethical reflections are to be made independent of any particular decision, as they are supposed to apply generically, meaning that they should be defined *a priori*, i.e. before the execution of the MCDA.

A posteriori, the pre-defined ethical principles can be applied to reflect on the results of the MCDA and the order of interventions in the league table. However, it should be noted that the general ethical principles policymakers wish to apply will already impact upon what they wish to include as an attribute in the MCDA or exclude from it. A typical example is age. When asked about therapeutic need, the majority of the population agreed that if you would have to develop a new intervention for a particular disease, you should first do it for the young patients with that disease. The underlying idea for this choice may be, as illustrated in literature, that the expected duration of benefit from the new treatment is longer in young than in old patients. Much less consensus exists on whether it is acceptable to reimburse a treatment to the young only if the treatment would have the same effect in all patients with a particular condition, independent of age. A recent study, in which Dutch, Belgian, Swedish and French decision makers were interviewed about the acceptability of discriminating in reimbursement based on age only, showed that this was generally not accepted and would in Sweden even be against the law.¹⁰²

The criterion of age is particular in several ways:

- Age cannot be modified by treatment, whereas all other attributes can.
- Age can be correlated with treatment effectiveness. For example, starting preventive dental care in early childhood has a proven better effectiveness than starting preventive dental care in the elderly. Effectiveness of treatment is a consideration in the reimbursement domain and thus decisions may vary by age, however, not *because of* variation in age but because of variation in effectiveness. The high importance of age in the results of the therapeutic need domain are in

a sense remarkable. All else equal, people favour the development of a new and better treatment for a young person over that for an old person. Losing 5 years of remaining life is considered worse for young people than for older people, even though for the elderly this means a much smaller number of remaining life years than for the young, who will only lose these years at the end of their normal life span. If they would discount the life years lost, losing 5 years would mean less for a young patient than for an old patient. We believe this can be explained by the fact that people did not clearly distinguish between “dying immediately” and “dying 5 years earlier”, which is what we observed in the model coefficients. In this case, it seems obvious that people will choose the younger patients when they are asked for whom a new treatment should be developed first, because the younger patients are assumed to benefit longer from the better outcomes of the new treatment.

- Unlike the other criteria, age is not a characteristic of a disease or its current treatment. It is a characteristic of a patient, like skin or natural hair colour or nationality. How to score such a disease- and treatment independent criterion? Should a disease that occurs mainly in the young get the highest score and diseases occurring mainly in the elderly get the lowest score? Would diseases that occur in all age groups then get a medium score? Suppose they do. In that case, therapeutic need in diseases occurring in all age groups would turn out lower than therapeutic need in diseases occurring mainly in the young. But this is not consistent with the preferences of the population, because these suggest that if a disease affects young patients (who are included in “all age groups”), high priority should be given to investments in better interventions. The need is hence high because also young people are affected; this is independent of whether also old people are affected. Should the judgment of the therapeutic need then be split up into ‘the need for the young’ and ‘the need for the old’? It could. The implications should be carefully considered. It would imply that, when a better treatment is being developed, this treatment might be withheld from the elderly patients, just because they are old and therefore lower on the therapeutic need list. According to the fair innings argument, this could be fair, because all humans start being young and then get older, which is very specific to the criterion of age, not applicable to other criteria or to characteristics such as gender and skin colour. This life-cycle



perspective could be considered relevant and a reason to discriminate between young and old in reimbursement decisions, even if all other criteria are equal. These are ethical choices that need to be made, which were not addressed in this study.

- For all other criteria, there is a logical relationship with need (from the patient's point of view): all else equal, a patient with a lower quality of life has a higher need than a patient with a higher quality of life, a patient with a life-threatening disease has a higher need than a patient with a non-life-threatening disease; but for age there is no such logical relationship: an older patient does not have a lower perceived need than a younger patient *only because* he is older.

As age as such is not a separate variable in our model, it could become an argument to change the ranking of disease on the therapeutic need league table.

It should be clear that MCDA is not a formula that leads to “easy” yes/no decisions. It is only through the consistent use and consideration of the relevant questions with the relevant criteria and their relative weights, that the decision-making process can become more consistent. Consistency implies rationality, in the sense that decisions about what the budget allows are more in line with what people consider important, both for individual patients as for the society as a whole. Moreover, the MCDA allows more transparency in the process. The remit of the advisory committees remains the same. The committee members remain responsible for the appraisal of the new interventions on different criteria. The only difference will now be that the weights given to each of these criteria will be those of the general public and not those of the committee members. Committees will still, as before, discuss additional criteria that are not included in the MCDA framework presented in this study and will still have to formulate an advice based on their appraisal. However, we hope this framework can help as a kind of reference point from which helps to justify –at least partly- the advices towards to general public and to create, as such, a societal ground for the decisions made.

6.3 Scoring rules

The development of detailed scoring rules was outside the scope of the current project. Some guidance, in general terms, is presented in Table 50. Detailed scoring rules, including guidance on which scale to use and how to move from the evidence to the scores, will be developed in a next report. Practical constraints related to the design of the survey led to a limited number of levels for each attribute included in the DCE. In real life, more differentiation will be needed between levels. For example, ‘more’ will be insufficient, as decision makers will want to make a distinction between ‘slightly more’ and ‘much more’ and probably many levels in between. However, the link with the levels used in the survey should be maintained somehow to ensure that the results of the survey are used in the right way. It means that some kind of mapping will be needed between the desired number of levels and the levels included in the survey. This refers to step ii under “criteria” in Figure 3 on page 29.

Besides the mapping issue, the issue of variety in the amount of evidence available on each attribute, the presentation of the evidence and the uncertainty of the evidence can also complicate the scoring.



Table 50 – Scoring rules for the criteria

Domain	Decision criteria	Scoring rule
Therapeutic need	Discomfort of current treatment	A higher level of discomfort gets a higher score
	Quality of life with current treatment	A lower quality of life gets a higher score
	Life expectancy despite current treatment	A higher reduction in life expectancy due to the disease, despite the current treatment, gets a higher score
Societal need	Disease-related public expenditure per patient	A higher disease-related public expenditure per patient gets a higher score
	Prevalence of the disease	A higher prevalence of the disease gets a higher score
Added value of new treatment	Impact on quality of life	A higher impact on quality of life gets a higher score
	Impact on prevalence of the disease	A higher impact on prevalence gets a higher score
	Impact on life expectancy	A higher impact on life expectancy gets a higher score
	Impact on discomfort of treatment	A higher reduction in discomfort of treatment gets a higher score
	Impact on disease-related public expenditure per patient	A higher impact on the public expenditure per patients gets a higher score

Key points

- The MCDA framework upon which previous, current and future research is built, focusses on the appraisal of therapeutic need, societal need and added value.
- The current study provides weights for scores that reflect the performance of a disease (in case of therapeutic or societal need) or an intervention (in case of added value) on each of the criteria described within the three domains.
- The scoring rules yet need to be developed. Scoring will be the responsibility of the advisory or decision making bodies, complemented with external experts as needed. Scores are a translation of the scientific evidence with respect to a disease or intervention on a pre-determined scale.

- The result will be three weighted MCDA scores, which need to be considered together in the decision process according to a certain logic.
- Relevant considerations that go beyond the criteria included in the MCDA, need to be addressed and made explicit as part of the decision process to maintain the biggest advantage of applying MCDA, being transparency and consistency in the decision-making process.



7 GENERAL DISCUSSION

7.1 Issues in MCDA research

A recent literature review of MCDA applications to assess the value of health care interventions defined eight challenges for those who want to apply MCDA in decision making.¹¹ They help to explain the place and value of our study and will therefore be treated one by one.

7.1.1 Choice of the MCDA approach

The first issue is related to the MCDA approach used. We used a value measurement approach, which implies that most outcomes relevant to a coverage decision can be traded off, i.e. poor performance on one criterion can be compensated by a good performance on another. Marsh et al. (2014) highlight that there might be ethical issues making trade-offs unacceptable. This is particularly the case when non-tradable criteria (such as access to care) are included in the MCDA.¹¹ We believe that the criteria eventually included in our MCDA do not include such criteria, although we cannot be completely sure about this. All depends on the fundamental ethical values the Belgian society wishes to uphold. For example, age could be an example of a non-tradable criterion. We included age in our scenario description, but did not treat it as a separate criterion in our MCDA tool because we only included age to allow respondents to judge the relevance of the impact of a disease on life expectancy. Besides the fact that we did not intend to include age as a separate attribute from the beginning, we think there are several additional arguments for not doing this.

First, age is a characteristic of a patient and cannot be changed by a new or better treatment. Therefore, it cannot be put on the same level as the other criteria, which are all disease-related characteristics under current treatment that can be changed by a better treatment (“quality of life”, “life expectancy” and “discomfort of the treatment”). It seems illogical to take age as *such* into account in health care reimbursement decisions. This is something else as taking decisions that are based on other features (e.g. clinical effectiveness) that eventually turn out to be to the disadvantage of specific age groups. For example, when preventive dental care is reimbursed only for children and not for the elderly that is mainly because of differences in (cost-) effectiveness and not because of age *per se*.

Second, it is not surprising at all that people choose the younger patient group when they have to judge the therapeutic need in two groups, all else equal. But it is much less certain that people would agree with not reimbursing interventions to elderly that are reimbursed to young patients only because of their age. Based on thirteen interviews with decision makers from Belgium, France, the Netherlands, and Sweden, Franken et al. (2014) concluded that decision makers do not consider age as criterion in decision making and interviewees from Sweden even mentioned that it would be against the law to use age as a decision criterion.¹⁰² This holds especially for discriminating in reimbursement of one single product or service for the same indication and with the same clinical benefit and cost-effectiveness. It would not be acceptable to reimburse a product for the young but not for the old (or vice versa) based on age alone if all else is equal.

Third, the government has the moral obligation to take decisions that are in society's best interest. Fundamental ethical principles public authorities might want to maintain could be, for instance, the respect for human dignity, individual rights and liberty, procedural impartiality, or the principle of solidarity.^{6, 103} For criteria such as quality of life, life expectancy and treatment discomfort, there does not seem to be a clear moral guiding principle, in which case it makes sense to follow the majority's wish. For age (independent of clinical effectiveness of treatments in the elderly), it seems hard to defend the position that elderly should be discriminated against simply because they are old despite an equal clinical benefit.

The operationalization of fundamental ethical values and principles is not always straightforward, as illustrated by many scholars.^{22, 103-105} Moreover, there seems to be no set of core values that is common to all health care systems.¹⁰⁶ For example, a comparative study between the Netherlands, Oregon (Canada), the UK and Sweden, found that people in Sweden attach particular importance to human dignity and the rights of individuals when setting priorities in health care, while the Netherlands and Oregon emphasize the efficient use of resources.¹⁰⁶ Each system needs to determine for itself how to balance different values. This does not only apply to age but also to aspects such as own responsibility for health problems.

Another argument against including age as a separate decision criterion is pragmatic. Many diseases occur in all age groups. It is unclear what the scoring rules for age would then have to be. Giving an intermediate score or giving no score at all would both be wrong. Giving no score immediately



reduces the total weighted score for therapeutic need drastically, implying that a disease that occurs in all age groups would always figure at the bottom of the list of therapeutic need. Giving an intermediate score has a similar effect: diseases occurring in all age groups would always be considered as having a lower therapeutic need than diseases occurring in younger age groups, all else equal, even though this disease also occurs in young people who should get priority according to the public.

7.1.2 Double counting criteria

The second issue highlighted by Marsh et al. (2014) is the risk of double counting criteria if both effectiveness (or cost) and cost-effectiveness is included in the MCDA. The authors found examples of MCDA studies including both, but also studies that excluded cost-effectiveness as a criterion, or compared the MCDA score against cost in a healthcare production possibility frontier.¹¹ The latter is also what we propose to do in the MCDA framework. We did not include cost-effectiveness in our MCDA, although we did include public expenditures induced by the disease per patient in the societal need domain. It is, in our view, impossible to leave out cost as a criterion entirely, because the economic impact of a disease has a value impact on society because of the opportunity costs of the resources devoted to that disease. An opportunity cost implies a value loss elsewhere in a publicly financed sector.

The production possibility frontier approach is similar to the efficiency frontier approach used for cost-effectiveness analysis, be it more extensive because it takes other considerations than clinical effectiveness in terms of “QALYs gained” into account. While in health economics interventions would be ranked in order of increasing incremental cost-effectiveness ratio, in MCDA interventions would be ranked in order of increasing MCDA score. However, extrapolating the cost-effectiveness analysis approach to the MCDA approach presented in this study, would require the following steps:

- The three MCDA scores need to be consolidated in one way or another, to allow the construction of one single league table of interventions that takes therapeutic need, societal need and added value into account.
- Additional considerations that require the adaptation of the ranking should be applied to arrive at a final league table.

- Depending on the available resources, first the highest ranked interventions should be financed, followed by lower ranked interventions until the budget is exhausted. As such the societal value of healthcare would be maximised.

7.1.3 Gaps in evidence and bias

The third issue relates to gaps and biases in evidence and subsequent difficulties in scoring an intervention on the criteria for which the evidence is lacking.¹¹ This is a general issue that cannot be easily resolved. In case of little evidence, less priority could be given, or reimbursement conditions could be imposed. Ways to deal with this issue will be explored further in a future study.

7.1.4 Inter-rater consistency

The fourth issue relates to the consistency in scoring of diseases or interventions on each of the criteria included in the MCDA: different raters might have different levels of understanding of the data and interpret scales differently, and the complexity of scales increases with the number of points on the scale.¹¹ We fully acknowledge this point and therefore recommend the development of a scoring guidance. Commission members, relevant stakeholders and experts will need to be involved in the guidance development process and afterwards be educated to ensure all raters have a similar understanding of the rating scales.

7.1.5 Choice of the weighting technique

The fifth observation of Marsh et al. (2014)¹¹ is that more debate and guidance is required on which weighting technique is appropriate under which circumstances. It was also our experience during this study that there are still many open questions with respect to the different techniques to elicit criteria weights. Although DCEs are frequently used in literature, we could identify only very few studies explaining how to derive level-independent criteria weights on a 0-1 scale from DCE results. For example, Lancsar et al. (2007)⁹⁵ compares four different methods for deriving attribute weights in health preferences: the log-likelihood-based method, the marginal rates of substitution, the Hicksian welfare measure, and the probability based method. They concluded that for all methods both the relative importance and the subsequent order of the attributes are similar, but not equal. Except



for the log-likelihood based method, all methods require either an attribute used as a common base for the relative importance or require a quantitative attribute (e.g. in monetary units, life years), something we did not have in all domains. Yet another method, proposed by an expert involved in our study, is to use the range of the estimated model coefficients per attribute to estimate the attributes' relative importance.

We eventually chose to use two methods to calculate relative weights: the log-likelihood method and the coefficient range method. As we only use discrete attributes, both methods are applicable to our DCE results.

The log-likelihood method has some advantages over the coefficient range method:

- the possibility to explicitly test statistical significance of the resulting weights;
- no risk of confounding relative importance (attribute weight) with the position of the attribute levels on the assumed underlying utility scales. The parameter estimates reflect both the relative importance of the attribute *and* the distance between attribute levels on the underlying utility scales. Consequently, when the underlying utility scales of the different attributes do not have a similar scale unit, there is a risk that this influences the parameter estimate size alongside the relative importance. The coefficient range method does not allow to disentangle these two aspects (relative importance and scale of underlying utility function).

In our study, the results for both methods are very similar, except for the Societal need domain. As there is no gold standard for measuring attribute weights, it is impossible to conclude which technique offers the most valid results. This is one of the reasons we chose to test two methods. Nevertheless, testing more alternative techniques could give indications about the robustness of the results presented in the current study.

A possible limitation of both methods is their relative dependence on the number of levels per attribute. Nevertheless, changing the number of levels of attributes only slightly changed the actual weights, but not the order of the attributes.

Moreover, increasing the number of levels per attribute is likely to make the DCE approach unfeasible, as it would require an increase in the number of choice sets per person. Based on our pilot survey, we considered the number of choice sets included in our survey (9) to be the maximum feasible to ensure an acceptable response rate, although in the literature it is found that the number of choice sets presented to respondents in empirical studies has increased from on average 12 in the period 1990-2000 to 14 in the period 2001-2008.

An additional concern is the feasibility of performing recurrent DCEs in the general public. Supposing that the weights people attach to different criteria change over time (which is very likely), one would have to repeat the research endeavour to collect and analyse the data. This is not the case when weights are elicited from the commission members. However, it has been explained before that the incorporation of public preferences is important for legitimacy reasons.

7.1.6 *Uncertainty in evidence*

The sixth issue relates to the quantifying of uncertainty.¹¹ How to deal with uncertainty about the evidence that should support the scoring of the criteria? Some MCDA studies included uncertainty as a separate criterion. There are different types of uncertainty: uncertainty about the expected outcomes and uncertainty about the chances of success. We considered to include uncertainty about expected outcomes in our DCE, but refrained from doing so because it would require an additional criterion for every other criterion already included because the level of uncertainty could in principle be different across criteria (e.g. we might be very uncertain about the effect of a new treatment on life expectancy because the follow-up in the available RCTs is too short to draw conclusions about the impact on life expectancy, but at the same time we might be quite certain about the impact of a new treatment on the comfort of treatment for the patient). We did not include uncertainty related to the probability of success because it has been demonstrated that the results for this criterion are highly subject to framing effects, and influenced by risk averseness.

Uncertainty could be taken into consideration after the application of the MCDA, i.e. when considering additional elements or criteria that would justify an uplift or downgrade of the ranking of an intervention. High uncertainty could thus be an argument for downgrading an intervention.



7.1.7 Interpretation of MCDA scores

The seventh issue relates to the interpretation of the MCDA output. In itself, the weighted average number generated by the MCDA is meaningless, aside from its use to rank interventions.¹¹ This contrasts with the output of the DCE that we used to determine the relative weights of the criteria. The multinomial logit model gives the probability that the general public would prefer a specific scenario out of the full set of possible scenarios. However, the DCE as conceived in our study cannot be used in place of the MCDA. We did not include a quantitative monetary attribute in our DCE (e.g. price or cost). It has been argued that this is problematic for the derivation of utility scores, because different attributes might have a different underlying scale. This results in difficulties in distinguishing the importance of the overall weight of a given dimension from the importance of the given levels within a dimension.⁸⁹

Nevertheless, we deliberately avoided using explicit monetary or quantitative levels in our DCE scenarios. We are aware that different people might give a different meaning to “much discomfort” or “little discomfort”. We did not perform a qualitative analysis of the underlying meaning of the different attribute levels for respondents. For our purposes, it was more important that the qualifiers had a meaning to respondents than that they had *the same* meaning for all respondents. What matters for determining therapeutic need is that respondents know that *patients* (as “*experts by experience*”) describe their level of discomfort as high or low. The assumption is that respondents would have the same experience if they would have been a patient; i.e. respondents trust the judgment of the current patients. In the application of the MCDA, the scoring will always be done by the members of the appraisal committee (based on evidence), *not by the general public*. This is comparable to, for instance, the use of the EQ-5D to measure health-related quality of life, where *patients* describe their own health status by means of the five dimensions and three or five levels per dimension of the instrument, and subsequently a *public* value is attached these descriptions.

Thus, we assume that the general public trusts the scoring of the committee members. Once the appraisal committee has judged that the discomfort is high, based on scientific evidence, the weight assigned to discomfort by the general public becomes relevant, because the weight reflects to which

extent discomfort should be taken into account in the decision making process, whatever high discomfort implies exactly.

A possible critique to this approach could be that the implied rankings in the MCDA can in that case be different from what the citizens want, because citizens might have another idea about what is high discomfort than patients. However, we considered the experience of patients to be more relevant than the subjective idea of what is high discomfort of someone who is not in the particular situation. The question really is “If you were a patient, and you would consider your discomfort to be high, how important would you consider that aspect for judging your need for a better treatment”. Whether discomfort is high in concrete situations, is then something for patients to judge, and their judgment to be measured in a scientifically sound manner.

The same applies to “high public expenditure” or “low public expenditure”. Respondents might not be able to judge what a high or low public expenditure is, but they can assume that decision makers are able to make this judgment based on their knowledge and experience with regard to the healthcare budget. For example, some people might find €3000 per patient per year high, others might find it low and still other will not have a clue, because much depends on what the implications of reimbursing such a treatment are for their personal expenditures. In the MCDA application, it will be the appraisal committee members who judge whether the public expenditure is high or low and the population (current study) that provides the weight for public expenditure. Thus, the judgment of whether discomfort/public expenditure is high is an expert judgment, while the relative importance of high discomfort for therapeutic need or high public expenditure for societal need is a population judgement. The objective is to keep a close relationship between the evidence (e.g. what do studies that measured patients’ discomfort with current treatment show) and the scoring of the criterion ‘discomfort’ in the MCDA, and to keep the value judgment of how important high discomfort is in judging therapeutic need separate from this. For quality of life and impact on life expectancy we did include quantitative levels. However, these quantities were used for their qualitative meaning: everyone would consider a 2/10 to be a low quality of life, and everyone would consider a 8/10 a high quality of life.

Hence, the DCE was a technique to derive relative attribute weights, rather than a kind of MCDA exercise, because we wanted to keep the opportunity open that the appraisal committees could score the attributes given their



knowledge and expertise. This forced us to take a two-step approach, being to perform a DCE first to obtain input data (weights) for the MCDA tool.

7.1.8 Impact of MCDA

Finally, very little is known about the impact of MCDA on decisions.¹¹ The authors would like to see more studies about which variations in MCDA approaches will have an impact on decisions. We would argue, as many other researchers, that even if there would be no impact on the actual decisions, the application of MCDA has its own merits. Surveys asking decision makers for their opinion about MCDA were quite positive: it provides a systematic approach to decision making, facilitates knowledge transfer and improves decision makers' understanding of interventions, identifies data gaps, forces decision makers to think through all relevant factors and improves the transparency of decisions. Moreover, it allows to communicate the rationale of their decisions.¹¹

7.2 Assumptions and limitations

7.2.1 Belgian citizens are not mere QALY maximizers

The mere fact of doing the survey also implies that we assume that people are not only interested in maximising QALYs, but rather wish to take the level of need for a new treatment into account when allocating limited health care resources.³⁹ As a consequence, a highly cost-effective treatment^e – i.e. a treatment that would contribute to the maximisation of QALYs – can still be considered of low priority for public reimbursement when the therapeutic or societal need is low. With this survey we demonstrated that the Belgian general public are not QALY-maximisers. This confirms the conclusions from previous research, both in Belgium² and elsewhere (see 4.2.3.3) that pure QALY maximization is never applicable, as people always want to take other criteria than cost-per-QALY into account.

7.2.2 A multi-layer MCDA is more manageable and acceptable to policymakers than an all-in-one MCDA

In our conceptual framework and with our data, the MCDA process takes place at multiple layers: therapeutic need, societal need, and added value. This implied that the criteria to be included in the DCE needed to satisfy the MCDA requirements at the domain-level only and not across domains, as for instance in the EVIDEM framework. This has two major advantages:

- By creating a stepwise hierarchical decision-making process, the number of criteria per step in the process diminishes as compared to an all-encompassing one-step decision-making process. It makes the considerations more manageable from the cognitive point of view.
- Criteria may be relevant at different levels of the decision-making process. In a step-wise process, where criteria are considered at each step, this is not a problem. The levels themselves do not have to be independent (by definition they are not, as it is a hierarchical process). The independence requirement only applies within each level.
- With this multi-layered MCDA, the separate layers could be used for other purposes than reimbursement decisions. For example, the layers relating to therapeutic and societal need could be used to identify the level of unmet need in case of specific diseases. This is particularly useful when a system wants to move from a supply-driven system to a more demand- or needs-driven system.

However, it also has disadvantages. By separating prevalence from severity of the disease under current treatment, we do not know how relatively higher the need is if it concerns a highly prevalent mild disease versus a not so prevalent severe disease.

^e Cost-effective in the neo-classical sense, i.e. contributing to the achievement of a maximum number of QALYs with a given amount of resources.



7.2.3 *A reasonableness test of the rankings will have to be performed*

Our model contains only a selection of criteria that may possibly be relevant to reimbursement decisions. This makes our model vulnerable to criticisms.¹⁰⁷ However, there is a trade-off between the number of criteria included in a decision-making process and the transparency –and consequently legitimacy – of the decision-making process. Nevertheless, the outcomes (i.e. ranking of diseases and interventions) of the application of our MCDA model will need to be subjected to a reasonableness test. The challenge will be to identify the reasonableness standard against which the rankings can be compared. As the Oregon experience has shown, the omission of certain attributes (such as contagiousness) might induce a wish to alter the rankings.¹⁰⁸ This can be perfectly legitimate. The citizen laboratories executed by the King Baudouin Foundation might provide insights into which criteria need to be considered on top of the ones included in the tool when determining the ranking of diseases or interventions.

The tool will have to be revised when certain criteria systematically pop up as being relevant and not included in the current MCDA model. In that case, a new public consultation will be needed to find out how important these criteria are relative to the other criteria and new weights will have to be determined.

7.3 Who should be involved in the application of the MCDA tool?

An issue we did not cover in our study is who should be involved in the decision-making process that uses the MCDA tool. Could the current commissions be maintained as they are, or are some modifications needed? Belgium has a deliberation-driven system. This is a choice. The alternative is to have an assessment-driven system. The distinction is not black and white. In both systems, appraisal will happen, and incorporation of social values will be necessary, in some way or another. Both systems can make legitimate decisions, based on procedural grounds (transparency, relevance, revisability and enforcement).

Tenbenschel (2002) has called, what we know in Belgium as the RIZIV-INAMI Commissions (e.g. the Drug Reimbursement Committee), “mediating, interpreting bodies”.¹⁰⁷ He argues that these bodies are essential to the

enhancement of priority setting processes that aspire to rationality and legitimacy. The approach taken in our study is the technocratic approach to public involvement, which is different from but can be combined with a participatory democratic approach. A purely technocratic approach tries to measure public values accurately and integrate these into decision-making processes, while a purely participatory democratic approach tries to reach public consensus based on an open debate and deliberation, such as through citizen’s councils as in the UK. Experience has shown that citizen’s councils rarely produce concrete advice that can be applied directly to specific priority setting decisions.¹⁰⁷ When the councils report back to the mediating bodies, the latter interpret the information and use it in their decision-making processes, or not. The process of transformation of public input into substantive decisions is highly opaque and the council’s reasoning could always be challenged on numerous grounds.¹⁰⁷ We would argue that participatory democratic approaches could benefit from the input of more quantitatively measured public preferences. As such, the citizen’s councils also have basis to start from when having their discussions about a particular technology.

We would argue that the committees could be maintained as they are, but ask and use the input of experts external to the committees for the scoring of the criteria, especially when evidence is lacking or highly uncertain; for example patients for the scoring of quality of life with current treatment and treatment discomfort, or specialists in a specific disease area for the scoring of uncertain clinical benefits. Rules for the declaration of conflicts of interest should be specified.



Key points

Methodological issues

- Although for many respondents the age of patients was important for making a choice about the therapeutic need, we did not determine a weight for age as a criterion for health care reimbursement decisions, because:
 - it has never been our intention to treat age as a separate criterion;
 - unlike the other criteria, age is not a disease or treatment related characteristic and can, hence, not be modified by another treatment;
 - there is no empirical support for the acceptability of discriminating in reimbursement based on age alone, independent of differences in clinical effectiveness;
- Cost-effectiveness was not included in our study,
 - to avoid double counting with cost and effectiveness attributes, and
 - because we considered it difficult to explain the meaning and implications of cost-effectiveness levels in a written survey.
- Many attributes did not have quantitative levels in our discrete choice experiments. A qualitative analysis of how respondents interpreted the qualitative levels was not performed. It was not essential for our approach that respondents gave the same meaning to the qualitative levels.

Assumptions underlying our research

- People living in Belgium are not mere QALY maximizers but want other considerations to be taken into account. The results of our study confirm this.
- A multi-layer MCDA is more manageable and acceptable to policymakers than an all-in MCDA.

- The criteria included in the current study are covering the most relevant and important ones. This is a debatable assumption. Revision and update of the current study is needed when specific criteria, not included in the current framework, continue to come up as additionally relevant for reimbursement decision making.

Further research

- Further research is needed on the scoring of diseases and interventions on the different criteria and on ways to deal with uncertainty and gaps in evidence, inter-rater inconsistency, .
- Further research is needed on other ways to determine the attribute weights. The log-likelihood approach and the coefficient range method mostly gave the same order of relative importance of criteria, but given the scarcity of research in this domain, more methods should be tested.
- A suggestion for future research is to start with piloting the use of one piece of the MCDA framework, e.g. related to the domain of therapeutic need, and learn from this pilot to refine the MCDA approach for the other domains.



8 CONCLUSION

Experience in other countries has shown that there are no simple or technical solutions to reimbursement decision making.¹⁰⁶ Ethical principles and decision-making criteria could help policy makers to make decisions, but will be unable to suggest the right decision.⁶ The central aim of this study was not to develop a prescriptive tool to make decisions about the reimbursement of new products or interventions, but rather to inform the decision making process about the public preferences for reimbursement decision criteria. An explicit ethical framework or set of values is important, not because it results in decisions, but because such a framework helps to make clear the nature of the trade-offs that are made.¹⁰⁶ This ethical framework should come on top of the MCDA.

The way decisions are taken in health care reimbursement commissions will not necessarily change due to our study, but it can make the decision process better informed about what the public values, more transparent and increase commissions' capacity to explain their reasoning for a particular decision.

The current study provides the weights to be used in the MCDA tool. However, more is needed before the MCDA tool can be used in practice. First, scoring rules should be developed to allow the appraisal committees to translate the scientific evidence from the HTA report into scores that reflect the severity level of a disease on the "needs" criteria and the performance level of an intervention on the added value criteria. Second, these scoring rules should be tested, e.g. by applying the MCDA tool with scoring rules to a number of past reimbursement submissions. Current or past members of reimbursement commissions should be actively involved in these pilots, as well as patient representatives.

The scoring rules will first be developed and pilot tested for the domain of therapeutic and societal need. This will allow testing the MCDA for identifying the unmet medical needs in Belgium. The identification of unmet medical needs is particularly relevant in the actual policy context, given the recent law regarding the accessibility of health care in case of unmet medical needs.¹⁰⁹ If the application of the MCDA for therapeutic and societal need proves applicable, also the application for added value will be developed further.



■ REFERENCES

1. Slovic P, Fischhoff B, Lichtenstein S. Behavioural decision theory. *Annual Review of Psychology*. 1977;28:1-39.
2. le Polain M, Franken M, Koopmanschap M, Cleemput I. Drug reimbursement systems: international comparison and policy recommendations. Brussels: Belgian Health Care Knowledge Centre (KCE); 2010. Health Services Research (HSR) KCE Reports 147CD/2010/10.273/90
3. Statistics Belgium. Bezit en gebruik van computer, internet, e-commerce, e-government, soort verbinding, ... [Web page].2014 [cited 13/06/2014]. Available from: http://statbel.fgov.be/nl/modules/publications/statistiques/arbeidsmarkt_levensomstandigheden/indicatoren_t_i_c_aupres_des_menages_et_individus_2012.jsp
4. Clark S, Weale A. Social values in health priority setting: a conceptual framework. *J Health Organ Manag*. 2012;26(3):293-316.
5. Daniels N, Sabin J. Limits to health care: fair procedures, democratic deliberation, and the legitimacy problem for insurers. *Philos Public Aff*. 1997;26(4):303-50.
6. Jennings B. Health policy in a new key: setting democratic priorities. *J Soc Issues*. 1993;49(2):169-84.
7. Gibson JL, Martin DK, Singer PA. Priority setting for new technologies in medicine: a transdisciplinary study. *BMC Health Serv Res*. 2002;2(1):14.
8. Devlin N, Sussex J. Incorporating multiple criteria in HTA. *Methods and processes*. London: Office of Health Economics; 2011.
9. Baltussen R, Niessen L. Priority setting of health interventions: the need for multi-criteria decision analysis. *Cost Eff Resour Alloc*. 2006;4:14.
10. Klein R. Dimensions of rationing: who should do what? *BMJ*. 1993;307(6899):309-11.
11. 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(4):345-65.



12. Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, et al. Eliciting public preferences for healthcare: a systematic review of techniques. *Health Technol Assess*. 2001;5(5):1-186.
13. Department for Communities and Local Government. Multi-criteria analysis: a manual. London: 2009. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/191506/Mult-crisis_analysis_a_manual.pdf
14. Nord E, Street A, Richardson J, Kuhse H, Singer P. The significance of age and duration of effect in social evaluation of health care. *Health Care Anal*. 1996;4(2):103-11.
15. Perrin AJ, McFarland K. Social Theory and Public Opinion. *Annual Review of Sociology*. 2011;37(1):87-107.
16. Shiell A, Seymour J, Hawe P, Cameron S. Are preferences over health states complete? *Health Econ*. 2000;9(1):47-55.
17. Oliver A. Complete preferences over health states: a reply to the paper by Shiell et al. *Health Econ*. 2000;9(8):727-8.
18. Wiseman V. Aggregating public preferences for healthcare: putting theory into practice. *Appl Health Econ Health Policy*. 2004;3(3):171-9.
19. Cookson R, Dolan P. Public views on health care rationing: a group discussion study. *Health Policy*. 1999;49(1-2):63-74.
20. Kasemsup V, Schommer JC, Cline RR, Hadsall RS. Citizen's preferences regarding principles to guide health-care allocation decisions in Thailand. *Value Health*. 2008;11(7):1194-202.
21. Golan O, Hansen P, Kaplan G, Tal O. Health technology prioritization: which criteria for prioritizing new technologies and what are their relative weights? *Health Policy*. 2011;102(2-3):126-35.
22. Basson MD. Choosing among candidates for scarce medical resources. *J Med Philos*. 1979;4(3):313-33.
23. Green C. Investigating public preferences on 'severity of health' as a relevant condition for setting healthcare priorities. *Soc Sci Med*. 2009;68(12):2247-55.
24. Ubel PA. How stable are people's preferences for giving priority to severely ill patients? *Soc Sci Med*. 1999;49(7):895-903.
25. Gallego G, Taylor SJ, McNeill P, Brien JA. Public views on priority setting for high cost medications in public hospitals in Australia. *Health Expect*. 2007;10(3):224-35.
26. Johri M, Damschroder LJ, Zikmund-Fisher BJ, Kim SY, Ubel PA. Can a moral reasoning exercise improve response quality to surveys of healthcare priorities? *J Med Ethics*. 2009;35(1):57-64.
27. Lim MK, Bae EY, Choi SE, Lee EK, Lee TJ. Eliciting public preference for health-care resource allocation in South Korea. *Value Health*. 2012;15(1 Suppl):S91-4.
28. Anderson M, Richardson J, McKie J, Iezzi A, Khan M. The relevance of personal characteristics in health care rationing: what the Australian public thinks and why. *Am J Econ Sociol*. 2011;70(1):131-51.
29. Zweibel NR, Cassel CK, Karrison T. Public attitudes about the use of chronological age as a criterion for allocating health care resources. *Gerontologist*. 1993;33(1):74-80.
30. Bowling A. Health care rationing: the public's debate. *BMJ*. 1996;312(7032):670-4.
31. Lees A, Scott N, Scott SN, MacDonald S, Campbell C. Deciding how NHS money is spent: a survey of general public and medical views. *Health Expect*. 2002;5(1):47-54.
32. Schwappach DL, Strasmann TJ. "Quick and dirty numbers"? The reliability of a stated-preference technique for the measurement of preferences for resource allocation. *J Health Econ*. 2006;25(3):432-48.
33. Mason H, Baker R, Donaldson C. Understanding public preferences for prioritizing health care interventions in England: does the type of health gain matter? *J Health Serv Res Policy*. 2011;16(2):81-9.
34. Mak B, Woo J, Bowling A, Wong F, Chau PH. Health care prioritization in ageing societies: influence of age, education, health literacy and culture. *Health Policy*. 2011;100(2-3):219-33.
35. Bowling A, Jacobson B, Southgate L. Explorations in consultation of the public and health professionals on priority setting in an inner London health district. *Soc Sci Med*. 1993;37(7):851-7.



36. Hadorn DC. Setting health care priorities in Oregon. Cost-effectiveness meets the rule of rescue. *JAMA*. 1991;265(17):2218-25.
37. Shmueli A. Survival vs. quality of life: a study of the Israeli public priorities in medical care. *Soc Sci Med*. 1999;49(3):297-302.
38. Chinitz D, Meislin R, Alster-Grau I. Values, institutions and shifting policy paradigms: expansion of the Israeli National Health Insurance Basket of Services. *Health Policy*. 2009;90(1):37-44.
39. Mortimer D, Segal L. Is the value of a life or life-year saved context specific? Further evidence from a discrete choice experiment. *Cost Eff Resour Alloc*. 2008;6:8.
40. Nord E, Richardson J, Street A, Kuhse H, Singer P. Who cares about cost? Does economic analysis impose or reflect social values? *Health Policy*. 1995;34(2):79-94.
41. Quintal C. Aversion to geographic inequality and geographic variation in preferences in the context of healthcare. *Appl Health Econ Health Policy*. 2009;7(2):121-36.
42. Williams A. Intergenerational equity: an exploration of the 'fair innings' argument. *Health Econ*. 1997;6(2):117-32.
43. Dolan P, Tsuchiya A. Health priorities and public preferences: the relative importance of past health experience and future health prospects. *J Health Econ*. 2005;24(4):703-14.
44. Denier Y. On personal responsibility and the human right to healthcare. *Camb Q Healthc Ethics*. 2005;14(2):224-34.
45. Fowler FJ, Jr., Berwick DM, Roman A, Massagli MP. Measuring public priorities for insurable health care. *Med Care*. 1994;32(6):625-39.
46. Elchardus M, Te Braak P. Uw gezondheidszorg, uw mening telt! Onderzoek uitgevoerd in opdracht van het Rijksinstituut voor Ziekte- en Invaliditeitsverzekering (RIZIV) naar aanleiding van zijn gouden jubileum. Eindverslag. Brussels: RIZIV; 2014. Available from: <http://www.riziv.be/information/nl/studies/study70/index.htm>
47. Luyten J. Equity, Efficiency and Public Health. Studies in the Ethics and Economics of Vaccination Policy. [Doctoral thesis]. Antwerp: University of Leuven and University of Antwerp; 2014.
48. Edlin R, Tsuchiya A, Dolan P. Public preferences for responsibility versus public preferences for reducing inequalities. *Health Econ*. 2012;21(12):1416-26.
49. Tymstra T, Andela M. Opinions of Dutch physicians, nurses, and citizens on health care policy, rationing, and technology. *JAMA*. 1993;270(24):2995-9.
50. Diederich A, Swait J, Wirsik N. Citizen participation in patient prioritization policy decisions: an empirical and experimental study on patients' characteristics. *PLoS One*. 2012;7(5):e36824.
51. Linley WG, Hughes DA. Societal views on NICE, cancer drugs fund and value-based pricing criteria for prioritising medicines: a cross-sectional survey of 4118 adults in Great Britain. *Health Econ*. 2013;22(8):948-64.
52. Eisenberg D, Freed GL. Reassessing how society prioritizes the health of young people. *Health Aff (Millwood)*. 2007;26(2):345-54.
53. Vetter N, Lewis P, Farrow S, Charny M. Who would you choose to save? *Health Serv J*. 1989;99(5163):976-7.
54. Mossialos E, King D. Citizens and rationing: analysis of a European survey. *Health Policy*. 1999;49(1-2):75-135.
55. Tsuchiya A, Dolan P, Shaw R. Measuring people's preferences regarding ageism in health: some methodological issues and some fresh evidence. *Soc Sci Med*. 2003;57(4):687-96.
56. Green C, Gerard K. Exploring the social value of health-care interventions: a stated preference discrete choice experiment. *Health Econ*. 2009;18(8):951-76.
57. Tsuchiya A, Dolan P. Do NHS clinicians and members of the public share the same views about reducing inequalities in health? *Soc Sci Med*. 2007;64(12):2499-503.
58. Blacksher E, Rigby E, Espey C. Public values, health inequality, and alternative notions of a "fair" response. *J Health Polit Policy Law*. 2010;35(6):889-920.



59. Bryan S, Roberts T, Heginbotham C, McCallum A. QALY-maximisation and public preferences: results from a general population survey. *Health Econ.* 2002;11(8):679-93.
60. Stolk EA, Pickee SJ, Ament AHJA, Busschbach JJV. Equity in health care prioritisation: an empirical inquiry into social value. *Health Policy.* 2005;74(3):343-55.
61. Watson V, Carnon A, Ryan M, Cox D. Involving the public in priority setting: a case study using discrete choice experiments. *J Public Health (Oxf).* 2012;34(2):253-60.
62. Louviere JJ, Flynn TN. Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in australia. *Patient.* 2010;3(4):275-83.
63. Nord E. The relevance of health state after treatment in prioritising between different patients. *J Med Ethics.* 1993;19(1):37-42.
64. Phillips M. Can the electronic nose really sniff out lung cancer? *Am J Respir Crit Care Med.* 2005;172(8):1060; author reply -1.
65. Schwappach DLB. The equivalence of numbers: the social value of avoiding health decline: an experimental Web-based study. *BMC Med Inform Decis Mak.* 2002;2:3.
66. McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, et al. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technol Assess.* 2001;5(31):1-256.
67. Sayre-McCord G. Metaethics. In: Edward N. Zalta UN, Colin Allen, R. Lanier Anderson editor. *The Stanford Encyclopedia of Philosophy.* Summer 2014 edition ed. Stanford, CA 94305: Metaphysics Research Lab; Centre for the Study of Language and Information, Stanford University; 2014.
68. van Roojen M. Moral Cognitivism versus non-cognitivism. In: Edward N. Zalta UN, Colin Allen, R. Lanier Anderson editor. *The Stanford Encyclopedia of Philosophy.* Fall 2014 edition ed. Stanford, CA 94305: Metaphysics Research Lab; Centre for the Study of Language and Information, Stanford University; 2014.
69. Driver J. The history of utilitarianism. In: Edward N. Zalta UN, Colin Allen, R. Lanier Anderson editor. *The Stanford Encyclopedia of Philosophy.* Winter 2014 edition ed. Stanford, CA 94305: Metaphysics Research Lab; Centre for the Study of Language and Information, Stanford University; 2014.
70. Steup M. Epistemology. In: Edward N. Zalta UN, Colin Allen, R. Lanier Anderson editor. *The Stanford Encyclopedia of Philosophy.* Spring 2014 edition ed. Stanford, CA 94305: Metaphysics Research Lab; Centre for the Study of Language and Information, Stanford University; 2014.
71. Roberts T, Bryan S, Heginbotham C, McCallum A. Public involvement in health care priority setting: An economic perspective. *Health Expect.* 1999;2(4):235-44.
72. Swoyer C. Relativism. In: Edward N. Zalta UN, Colin Allen, R. Lanier Anderson editor. *The Stanford Encyclopedia of Philosophy.* Summer 2014 edition ed. Stanford, CA 94305: Metaphysics Research Lab; Centre for the Study of Language and Information, Stanford University; 2014.
73. Baron J, Ubel PA. Revising a priority list based on cost-effectiveness: the role of the prominence effect and distorted utility judgments. *Med Decis Making.* 2001;21(4):278-87.
74. Schwarzingler M, Lanoe JL, Nord E, Durand-Zaleski I. Lack of multiplicative transitivity in person trade-off responses. *Health Econ.* 2004;13(2):171-81.
75. Dolan P, Tsuchiya A. The person trade-off method and the transitivity principle: an example from preferences over age weighting. *Health Econ.* 2003;12(6):505-10.
76. Kinnunen J, Lammintakanen J, Myllykangas M, Ryyanen OP, Takala J. Health care priorities as a problem of local resource allocation. *Int J Health Plann Manage.* 1998;13(3):216-29.
77. Matschinger H, Angermeyer MC. The public's preferences concerning the allocation of financial resources to health care: results from a representative population survey in Germany. *Eur Psychiatry.* 2004;19(8):478-82.
78. Lee JA, Soutar GN, Louviere J. Measuring values using best-worst scaling: The LOV example. *Psychology and Marketing.* 2007;24(12):1043-58.



79. Danner M, Hummel JM, Volz F, van Manen JG, Wiegard B, Dintsios C-M, et al. Integrating patients' views into health technology assessment: Analytic hierarchy process (AHP) as a method to elicit patient preferences. *Int J Technol Assess Health Care*. 2011;27(4):369-75.
80. Hummel MJM, Volz F, van Manen JG, Danner M, Dintsios C-M, Ijzerman MJ, et al. Using the analytic hierarchy process to elicit patient preferences: prioritizing multiple outcome measures of antidepressant drug treatment. *Patient*. 2012;5(4):225-37.
81. Smith RD. Construction of the contingent valuation market in health care: a critical assessment. *Health Econ*. 2003;12(8):609-28.
82. Bryan S, Gold L, Sheldon R, Buxton M. Preference measurement using conjoint methods: an empirical investigation of reliability. *Health Econ*. 2000;9(5):385-95.
83. Whitty J, Ratcliffe J, Chen G, Schuffham P. A comparison of discrete choice and best worst scaling methods to assess Australian public preferences for the funding of new health technologies. In: Paper submitted for presentation at the International Choice Modelling Conference, 3-5 July 2013, Sydney Australia.; 2013. p. 32.
84. Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PF. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best-worst scaling? *Eur J Health Econ*. 2014;15(3):281-8.
85. Potoglou D, Burge P, Flynn T, Netten A, Malley J, Forder J, et al. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. *Soc Sci Med*. 2011;72(10):1717-27.
86. Severin F, Schmidtke J, Muhlbacher A, Rogowski WH. Eliciting preferences for priority setting in genetic testing: a pilot study comparing best-worst scaling and discrete-choice experiments. *Eur J Hum Genet*. 2013;21(11):1202-8.
87. Carton A, Statistics icwsmotaPa. Kwaliteitsrichtlijnen bij het uitvoeren van surveyonderzoek. In: Statistiek MvdVG-APe, editor. Brussel; 2001. p. 59 pp.
88. Billiet J, Carton A. Dataverzameling: gestandaardiseerde interviews en zelf-in-te-vullen vragenlijsten. In: H. BJW, editor. Een samenleving onderzocht: methoden van sociaal-wetenschappelijk onderzoek. Antwerpen: De Boeck; 2003. p. 285-314.
89. de Bekker-Grob EW, Hol L, Donkers B, van Dam L, Habbema JD, van Leerdam ME, et al. Labeled versus unlabeled discrete choice experiments in health economics: an application to colorectal cancer screening. *Value Health*. 2010;13(2):315-23.
90. Koninklijk Besluit van 21 december 2001 tot vaststelling van de procedures, termijnen en voorwaarden inzake de tegemoetkoming van de verplichte verzekering voor geneeskundige verzorging en uitkeringen in de kosten van farmaceutische specialiteiten, B.S. 29 december 2001. Arrêté royal de 21 décembre 2001 fixant les procédures, délais et conditions concernant l'intervention de l'assurance obligatoire soins de santé et indemnités dans le coût des spécialités pharmaceutiques, M.B. le 296 décembre 2001., 2001.
91. Reed Johnson F, Lancsar E, Marshall D, Kilambi V, Muhlbacher A, Regier DA, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value Health*. 2013;16(1):3-13.
92. Goos P, Jones B. Optimal Design of Experiments: A Case Study Approach. Wiley; 2011.
93. Kessels R, Goos P, Vandebroek M. A Comparison of Criteria to Design Efficient Choice Experiments. *Journal of Marketing Research*. 2006;43(3):409-19.
94. Kessels R, Jones B, Goos P. Bayesian optimal designs for discrete choice experiments with partial profiles. *Journal of Choice Modelling*. 2011;4(3):52-74.



95. Lancsar E, Louviere J, Flynn T. Several methods to investigate relative attribute impact in stated preference experiments. *Soc Sci Med.* 2007;64(8):1738-53.
96. Croissant Y. mlogit: multinomial logit model (R package version 0.2-4). In; 2013.
97. Hensher DA, Rose JM, Greene WH. *Applied Choice Analysis: A Primer.* Cambridge University Press; 2005.
98. Bech M, Gyrd-Hansen D. Effects coding in discrete choice experiments. *Health Econ.* 2005;14(10):1079-83.
99. European Commission. Europeans and their languages. 2006. Special Eurobarometer 243
100. Tafforeau J. Subjectieve gezondheid. In: Van der Heyden J, Charafeddine R, editors. *Gezondheidsenquête 2013. Rapport 1: Gezondheid en Welzijn.* Brussel: WIV-ISP; 2014.
101. Commission E. *The European Union Labour Force Survey (EU LFS).* 2013.
102. Franken M. *Decision Making in Drug Reimbursement.* Rotterdam (the Netherlands): Erasmus University Rotterdam; 2014.
103. Asplund K. Ethical issues in healthcare prioritization: a medical viewpoint. *Br J Urol.* 1995;76 Suppl 2:49-54.
104. Biron L, Rumbold B, Faden R. Social value judgments in healthcare: a philosophical critique. *J Health Organ Manag.* 2012;26(3):317-30.
105. Daniels N. The Articulation of Values and Principles Involved in Health Care Reform. *J Med Philos.* 1994;19(5):425-33.
106. Ham C. Priority setting in health care: learning from international experience. *Health Policy.* 1997;42(1):49-66.
107. Tenbensen T. Interpreting public input into priority-setting: the role of mediating institutions. *Health Policy.* 2002;62(2):173-94.
108. Brannigan M. Oregon's experiment. *Health Care Anal.* 1993;1(1):15-32.
109. Wet van 7 februari 2014 houdende diverse bepalingen inzake de toegankelijkheid van de gezondheidszorg / Loi de 7 février 2014 portant des dispositions diverses en matière d'accessibilité aux soins de santé. In: *Belgisch Staatsblad / Moniteur Belge* de 25/02/2014; 2014.

