# Measurement of student preferences regarding criteria used in drug reimbursement decisions

Author: Lucas van Schaik
Student number: 585816
Supervisor: S. Himmler
Date: 23-06-2021
Location: Rotterdam
Wordcount: 10,929

Erasmus University Rotterdam

# Abstract

**Objectives**

The study aims to elicit preferences of Dutch students for criteria used and potential additional relevant criteria in the drug reimbursement process with a discrete choice experiment (DCE). In addition, differences in these preferences between students with a medical and another educational background are estimated.

**Methods**

A DCE with a D-efficient design was constructed and held among Dutch students. They had to select their preferred alternative among 2 scenarios in 12 choice sets. Each scenario consisted of 6 attributes; disease severity, health improvement, composition of health gain, cost-effectiveness, size of the patient population and the age of the patient population. Each attribute contained three different levels. Demographic information including educational background was obtained. The results were analysed with a mixed logit model. Interactions for students with a medical educational background compared to students with another educational background were added to the model to estimate differences in preferences regarding criteria used in the reimbursement process between students with different backgrounds.

**Results**

115 students completed the DCE and significantly ranked age of the patient population, cost-effectiveness, health improvement and disease severity as the most important attributes. The attribute population size was also significant. However, the composition of health gain was not. Students preferred treatments aimed at younger people, treatments with a good cost-effectiveness, treatments with a large health improvement and treatments aimed at people with high disease severities. No statistical differences in preferences were found between students with a medical and other educational backgrounds.

**Conclusion**

Students ranked both currently and not currently used criteria as important for the reimbursement process of drugs. If the preference for treatments aimed at younger people is confirmed by research in other age groups, age adjustments could be a way to incorporate this criterion into the reimbursement process. The criteria that are currently used in the Dutch reimbursement process were ranked as important by students, which indicates alignment of the reimbursement process with preferences from the public. DCEs are an excellent tool for eliciting preferences in the health care system and provide valuable information for the development of health technology assessment guidelines.

# Table of contents

# Chapter 1. Introduction

## 1.1 Problem definition and rationale

Health care costs in the Netherlands are expected to increase until at least the year 2060. With unchanged policy in this area, costs will rise by approximately 2.8% each year (1). One study identified population size, population age, service price and service intensity as drivers for the increasing health care costs (2). Extramural and intramural drug costs are rising even faster compared to the total health care costs. In 2019, the costs of extramural drugs increased by 4.7% to a total of 4.9 billion euros. Expensive intramural drugs increased by 6.4% to 2.4 billion euros in the same period (3). One reason for the rapidly increasing pharmaceutical costs is the shift from blockbuster small molecules towards biologicals and other breakthrough treatments. These new agents have higher prices compared to small molecules and will most likely have a substantial impact on the overall health care costs (4).

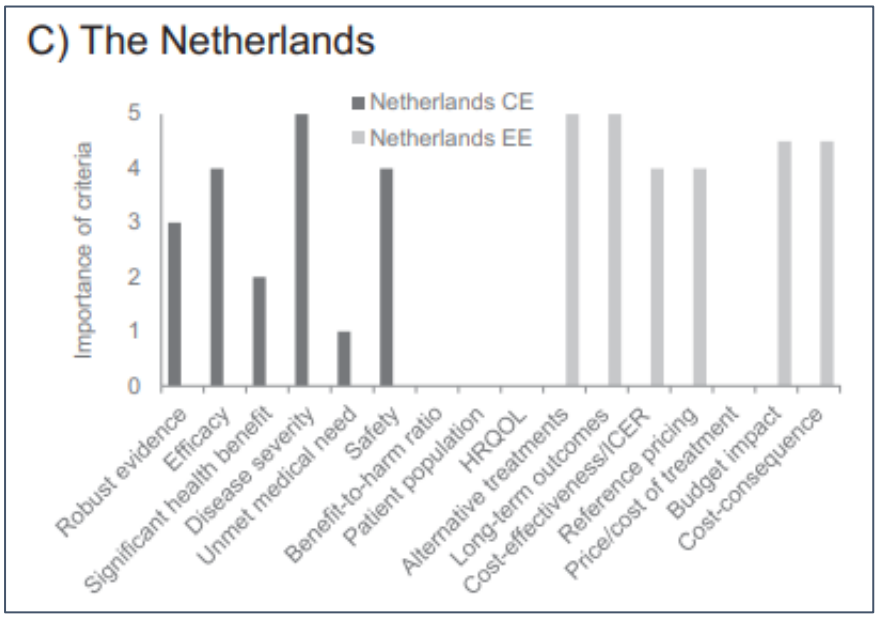
Growing concerns about high medical costs fuelled the introduction of health technology assessment (HTA) around 1970. HTA can be described as the evaluation of a health intervention by reviewing or producing evidence. This evaluation can be used to identify and eliminate interventions that are not safe, not effective or not cost-effective. HTA advanced to evaluating different health technologies, ranging from drugs to medical devices and surgical procedures (5). HTA analyses are mostly done in the form of a cost-effectiveness analysis (CEA) or a cost-utility analysis (CUA). The difference between these two forms is the outcome that is used for the analysis. A CEA often measures health effects as life-years gained, whereas a CUA also takes quality of life (QoL) into account and uses the measure quality-adjusted-life-years (QALY). Both analyses use these health effects to calculate the incremental cost-effectiveness ratio (ICER) (6,7). The current method of HTA for drugs in the Netherlands will be discussed in detail in chapter 2.1 ‘Institutional background’.

Historically, cost-effectiveness or cost-utility has been the focus of HTA. Other criteria are considered in HTA, but only informal. Schmitz et al. found that these other criteria besides cost-effectiveness or cost-utility influenced HTA decisions in the past (8). Informal use of criteria reduces the transparency of HTA. More recently, an increasing interest has been seen in the formal incorporation of these other criteria. The multi-criteria decision analysis approach was developed to achieve this goal. This method is a way of formally including different criteria, resulting in more transparency and consistency. When implementing such an approach, it is important to identify all possible available decision criteria and to determine the relative importance of these criteria (8).

Akehurst et al. provided an overview of the different HTA bodies in Europe and asked HTA experts from different countries to assess the importance of a set of decision-making criteria. The results for the Dutch HTA experts are shown in figure 1. The clinical expert ranked disease severity, efficacy and safety as the most important criteria. The economical expert ranked the availability of alternative

treatments, long-term outcomes, budget impact and the cost-consequence the most important criteria. The selected most important criteria varied between experts from different countries. For example, clinical and economical experts from the United Kingdom ranked cost-effectiveness as the most important criterion, whereas both experts from France thought disease severity the most important. Experts from Germany and France placed very little to no importance on the economic criteria, which differed from experts from other countries (9).

*Figure 1. Most influential decision criteria for HTA according to experts from the Netherlands* (9).




*CE Clinical expert, EE Economical expert, HRQOL health-related quality of life; ICER, incremental cost-effectiveness ratio.*

Criteria used in HTA should not only reflect the preferences of HTA-experts, but should also reflect the preferences of patients. That was one of the key findings of an extensive review of multiple HTA systems in Europe performed by Sorenson et al (5). Currently, preferences are taken into account in the form of safety and effectiveness measures. However, these measures do not capture broader patient preferences, for example, the preference for a certain treatment or acceptability of certain side effects (5). Patients or consumers of healthcare may value other outcomes of healthcare besides the QALYs that are often used in a CUA (6). Public preferences are important for HTA bodies and should be taken into account when creating HTA frameworks. Towards this end, the National Institute for Health and Care Excellence (NICE) established a Citizens' Council to collect society's preferences to incorporate in the development of HTA guidelines (5). HTA bodies want their decisions to align with the needs and preferences of the patients (10). Deviation of these preferences will eventually lead to suboptimal decision-making because it will lead to a lower utility for the patients.

Mülbacher and Juhnke reviewed the differences and similarities of patient preferences and physician's preferences. The included studies were heterogeneous in their conclusions. Studies were identified that

showed no meaningful differences, but the majority of the included studies demonstrated poor concordance between patient preferences and physician judgment (11). Although this research does not examine HTA preferences, it does show that preferences between different healthcare stakeholders can differ. It is therefore important to elicit the preferences of both patients and healthcare experts and to establish if there are differences in preferences between these groups.

One way of measuring preferences is with discrete choice experiments (DCE). Usually, DCEs consists of several choice sets where respondents have to choose between two or more hypothetical alternatives. DCEs could be useful for eliciting patient and expert preferences towards criteria that are potentially relevant or used in HTA (12).

The objective of this study is to elicit preferences of the society for criteria used and potential additional relevant criteria in the drug reimbursement process with a DCE. The main research question will be the following: What are the preferences of students regarding different HTA criteria in the Netherlands? This also entails examining the following sub-questions :

1. Are the criteria that are currently used in the reimbursement process more important than additional criteria that are not currently used in the reimbursement process (e.g. number of patients)?

2. Is there heterogeneity regarding preferences between individuals with a medical background compared with individuals with other backgrounds towards these criteria?

Sub-question two is an attempt to capture potential differences between two groups representing the general public and healthcare professionals. Although the respondents are still students, they will eventually become healthcare professionals and see the healthcare system with a different perspective compared to the students without a medical background.

## 1.2 Reader guide

The ‘Background’ section contains information about the current HTA process performed by Zorginstituut Nederland (ZIN), the theoretical framework and identified literature discussing DCEs in general and earlier results on preferences of HTA criteria. The attributes and level identification and selection are extensively discussed in the ‘Methods’ section. This section also contains information about the experimental and the survey design, the data collection, the sample size calculation and the statistical analysis. The ‘Results’ section first presents the demographic information and the evaluation results. After that, this section discusses the model selection and contains the results of the main model. The results are discussed and placed in perspective in the ‘Discussion and conclusion’ section. The strengths and limitations of the study are described in this part as well.

# Chapter 2. Background

## 2.1 Institutional background

In the Netherlands the current drug reimbursement process is performed by ZIN. This institute advises the minister of Health, Wellbeing and Sport (VWS) which drugs to approve for reimbursement. Subsequently, the minister of VWS chooses to approve or reject the drug for reimbursement, or chooses to negotiate with the pharmaceutical company to lower the price of the drug (13). To assess which drugs should receive a reimbursement status, ZIN asks the following questions:

1. Is there an important health problem?
2. Is there an available treatment that can solve this problem?
3. Are the effects of the treatment in a reasonable relation with the costs of the treatment?
4. Are the costs of the treatment outside the scope of the patient, but inside the scope of the society (13)?

The current extramural drug reimbursement process of ZIN consists of a pharmaco-therapeutic analysis and a pharmaco-economic analysis. In the pharmaco-therapeutic analysis, the added benefit of a new drug is evaluated. Favourable and unfavourable effects of the treatment are assessed and compared to the standard of care. When the effects are similar between a new drug and the standard of care, other criteria such as applicability, experience and ease of use can be included in the assessment. Based on this evaluation, new treatments are divided into three categories: treatments with a lower therapeutic value compared to other available treatments, treatments with an equal therapeutic value compared to other available treatments and treatments with a higher therapeutic value compared to other available treatments. After this therapeutic assessment, an economic assessment takes place (14).

The pharmaco-economic analysis performed by ZIN uses a societal perspective, which includes all relevant costs that are made by society. Subsequently, the patient population, intervention, control group, outcome and the time horizon are determined. There are different methods available to perform a pharmaco-economic analysis. ZIN chooses to use the CUA method and if necessary, a budget impact analysis. Future effects and costs are discounted at respectively 1.5% and 4.0%. The uncertainty in the model is determined by sensitivity analyses. The inputs for the pharmaco-economic analysis are the effects in QALYs and the associated costs. With these data, the ICER is calculated. (15) There are three thresholds for the calculated ICER and they are based on the disease severity. Table 1 shows the different threshold values for the different health states. The disease severity is measured as the proportional shortfall of the future life years and QoL of a person with a certain disease compared to a person without that disease. ZIN is willing to accept higher costs for higher disease severities. (13)

*Table 1. Incremental cost-effectiveness ratio (ICER) thresholds for different disease severities. Diseases with a disease severity < 0.1 are generally not approved for reimbursement* (13).

| Disease severity | Maximum ICER |
|---|---|
| 0.1 ≤ 0.4 | Up to €20,000 |
| 0.41 ≤ 0.70 | Up to €50,000 |
| 0.71 ≤ 1.0 | Up to €80,000 |

*ICER incremental cost-effectiveness ratio.*

## 2.2 Theoretical foundation of discrete choice experiments

DCEs are used to obtain choice preferences. According to economists, choices are based on an underlying choice process which assumes that individuals will choose the option that results in the highest utility. These choice preferences can be obtained by actual choices (revealed preferences) or by asking individuals to choose between hypothetical scenarios (stated preferences), which happens during a DCE (6). The theoretical framework of both methods of data collection is consistent with Lancaster's theory of value, which assumes that consumers derive the utility of a good from the good's characteristics, referred to as attributes. It is therefore possible to deconstruct an object, or drug in this particular case, in multiple attributes and different variants of the attributes, referred to as levels. The combination of the different levels of the attributes of an object results in the total utility of the object (16).

A DCE contains a set of choice tasks with hypothetical alternatives from which respondents have to choose their preferred alternative. Respondents make implicit trade-offs each time they complete a choice task and preferences can be obtained from these responses. McFadden developed the theoretical foundation of DCEs, the random utility theory (RUT) (17), drawing on the previous work of Thurstone (18). RUT assumes that the utility (U) of an individual (i) is based on the attributes of the alternative (j). This utility can be divided into a systematic and a random component, written down in equation *I* where $V_{ij}$ is the systematic component and $\varepsilon_{ij}$ is the random component (12).

$$U_{ij} = V_{ij} + \varepsilon_{ij} \qquad I$$

The systematic component depends on the attributes of the alternative and on the attributes of the individual. If the vector of the attributes of the alternatives equals *x* and the vector of the characteristics of the individuals equals z, equation *II* can be made. The vectors $\beta$ and $\delta$ in the equation represent the influence or coefficients of the alternatives and the individual (6).

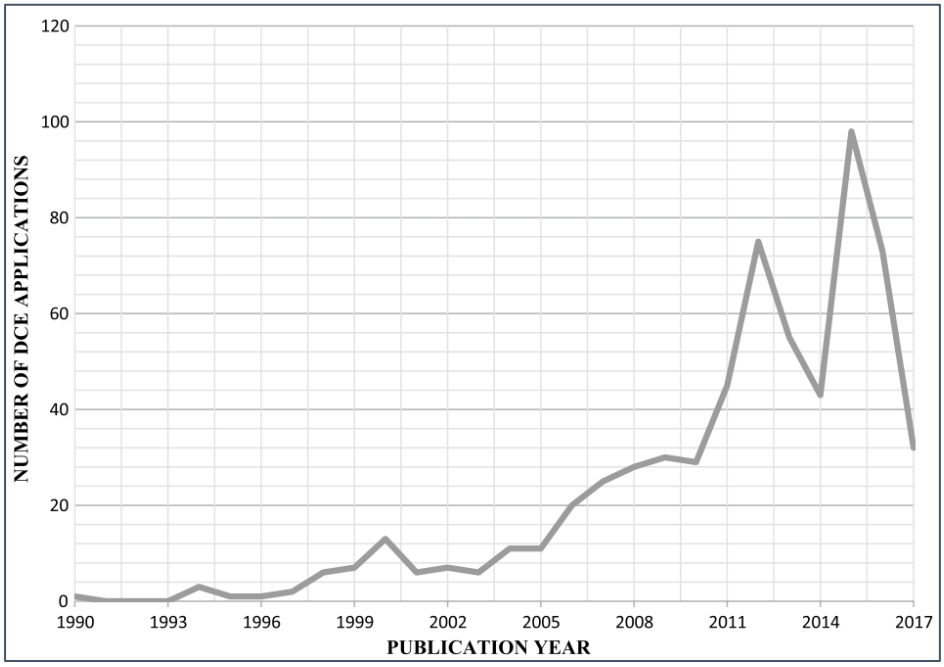
$$V_{ij} = \beta' x_{ij} + \delta' z_{ij} \qquad II$$

In this study, the vector $\beta$ will be estimated for all the attributes. A DCE is ideal for preference elicitation because this vector can be used to show the importance that respondents place on the different attributes.

## 2.3 Literature on preferences towards HTA criteria

An increasing amount of published DCEs has been seen in the field of health economics. This trend is depicted in figure 2 and is not very consistent from year to year. It is thought that this lack of consistency might be a result of other competing stated preference methods . The increase in DCEs is accompanied by a trend towards the use of more sophisticated econometric models and more sophisticated software, for example, the use of Ngene to generate D-efficient designs (19). The use of fractional designs and a qualitative approach for level selection is also becoming more common in DCEs. External validity tests are still lacking in DCEs, as only 2% of the identified studies by Soekhai et al. reported external validity. The lack of external validity tests could be due to the difficulties involved with the external validation of DCEs. Internal validity is tested more often, mostly with tests for theoretical and face validity and consistency (19).

*Figure 2. The number of discrete choice experiments in the area of health economics by publication year* (19).




A systematic review by Trapero-Bertran et al. identified attributes that were used in DCEs for eliciting preferences towards different criteria that are potentially relevant or used in HTA. More than half of the included papers were aimed at drug treatments. The attributes improvement in health, side effects and cost of treatment were most prevalent. Waiting time for the treatment, severity of the disease and value for money were less relevant, but all attributes can be considered to capture and describe societal preferences concerning HTA (20).

The following paragraphs summarize the most important results of published preference elicitation towards different criteria that could be relevant for the HTA of drugs.

Koopmanschap et al. performed a DCE among Dutch healthcare professionals and advanced HTA students. 66 Respondents were recruited and analysed with a Multinominal logit model. The DCE consisted out of six attributes: budget impact, national savings in costs of absence from work per year, disease severity, ICER, QALY's gained, composition of health gain and the uncertainty in the costs per QALY. The most important criteria were the severity of disease, the ICER, individual health gain and the budget impact. Subgroup analyses showed that especially the HTA students preferred QoL improvement over extension of life (EoL). EoL might be discounted more compared to QoL improvement because EoL is only beneficial in the distant future for students (21).

Another study explored the preferences of the members of NICE with a DCE. Attributes investigated were the ICER, uncertainty, age, disease severity and the availability of other therapies. The findings of the DCE were in line with the guidelines of NICE, as the age of the treatment population was not significant in the study. Other criteria were all significant and the respondents were unlikely to reimburse technologies with bad economic profiles, unless there were no alternative treatments present or when the disease severity was high (22).

In contrast to the aforementioned study which elicited preferences of NICE members only, Wranik et al. revealed preferences of HTA stakeholders from multiple countries (23). The included attributes were survival benefit, added cost per patient, number of patients, other treatment options and adverse events. The conducted DCE had a single scenario design to imitate the single scenario choice setting from real-world HTA questions. The respondents were asked whether they would reimburse the drug and how difficult the choice was. In addition, respondents had to perform a best-worse scaling. Clinical benefit was ranked the most important characteristic by the HTA policymakers. When other treatment options were not available or when the target population was large, respondents were willing to settle for lower clinical benefit. The unmet need played a role in these decisions, as people seem to settle for less when there is a higher unmet need. Larger populations in this case might be interpreted as a greater need in society (23).

In addition to the studies that showed preferences of HTA stakeholders, a DCE held in the general public was conducted by Green and Gerard. NICE states that their advice regarding health technologies should be in line with values that are held by the population (10). Preferences of the general public are therefore meaningful to HTA policies and should be evaluated. Four attributes were included in Green and Gerard's DCE: disease severity, health improvement, value for money (cost-effectiveness) and availability of other treatments. The general public ranked the level of health improvement the most important, followed by value for money. Disease severity ranked third highest and the attribute availability of other treatments ranked least important and was the only non-significant attribute (24).

# Chapter 3. Research methods

## 3.1 Identification of attributes and levels

A DCE was developed in this study to answer the research questions. Attribute and level selection is an important part of the DCE design, as poor selection leads to invalid results. The number of attributes was limited to 7 because a higher number of attributes increases choice task complexity and decreases behavioural efficiency. The behavioural efficiency of respondents entails choice consistency, respondent fatigue, drop-out rate and using simple heuristics (25,26). The selection of the used attributes and levels in the DCE was based on a literature search, the criteria used in the current drug reimbursement process by ZIN and a pilot survey conducted among five students with various educational backgrounds. The pilot was held to see whether the attributes and levels were understandable and to find exploratory insights in the relative importance of the attributes. Seven possible attributes were identified and included in the pilot survey, which is listed in appendix 1. Not all included attributes are currently used in the drug reimbursement process. This is done to elicit if other criteria could be a potential addition to the current reimbursement process, for example by adding these criteria to the aforementioned multi-criteria decision analysis method.

The target population of the DCE consisted of students with various backgrounds and no required knowledge in the area of HTA. This resulted in an attribute and level selection that needed to be understandable for the general public by avoiding difficult concepts. To improve the understandability of the attributes, respondents received a short explanation containing relevant information concerning the attributes and levels. This explanation is listed in appendix 2.

No clear best practice is known for level selection. The number of levels for each attribute should be limited to three to four. It is important to avoid extreme levels and to test the levels in a pilot if possible (27). The level selection was also focused on improving the understandability of the DCE. A numerical price attribute can be used to calculate willingness to pay and provides valuable information about the trade-offs respondents make (6). However, costs of drugs or ICERs could be substantial and might confuse respondents without experience in the HTA field. Therefore, it was chosen to not include numerical price levels, but to keep levels general as Green and Gerard did in their DCE (24). General levels were not only chosen for the price attribute, but for all attributes to make sure that the cognitive burden was not too high for the respondents, which improves the behavioural efficiency.

Table 2 shows the final selection of attributes and levels. The reasoning behind each of the selected attributes and levels will be discussed below.

*Disease severity*

Disease severity is a criterion currently used to determine the thresholds for the ICER in the drug reimbursement process in the Netherlands. It reflects the equity preference of the society, as ZIN finds it acceptable to spend more money on people with a higher disease burden (13). It is therefore an extremely important criterion in the HTA process. This is reflected in the literature, where disease severity is often included in the DCEs and ranked as the most important attribute in several DCEs (20,21,24). The importance of this attribute was also shown in the results of the conducted pilot as disease severity was ranked the most important attribute.

The disease severity contained three levels: a small impediment, moderate impediment and large impediment. These three severities were linked to a QoL score and examples in the explanation section of the survey. Hypertension was an example of a low disease severity. Chronic pulmonary disease (COPD) and a severe form of Alzheimer were examples of a moderate and high disease severity, respectively. Disease severity was framed as 'impediment' in the choice tasks of the DCE to make sure that the respondents associated a higher impediment with a higher burden of disease.

*Health improvement*

The efficacy of a drug was captured in this attribute. The health improvement is determined in the pharmaco-therapeutic analysis of the drug reimbursement process and is therefore an important criterion in HTA decisions. This attribute was the most frequently included in HTA related DCEs (20). Not surprisingly, this attribute also ranked fairly high in the conducted pilot and was therefore an obvious inclusion in the DCE. The different levels of health improvement were a low, moderate and high health improvement.

*Cost-effectiveness*

Cost-effectiveness is also an essential attribute in the current drug reimbursement process. Cost-effectiveness reflects the outcome of the HTA process, where the effects of a drug are compared to the costs of the drug (13). Cost attributes were present in the literature as total costs, cost-effectiveness or budget impact (20). Cost-effectiveness was also included in DCEs performed in a similar institutional setting (21) and in the general public, although named value for money in this particular DCE (24). This confirmed that the attribute is relevant in the setting of the Dutch healthcare system and understandable enough for the target population. The pilot study showed that people preferred the term 'cost-effectiveness' over 'value for money' and that the attribute was understandable.

Cost-effectiveness was chosen over other cost attributes, such as drug costs and budget impact. Drug costs were not used in the place of cost-effectiveness because with this method people would have to make their own implicit calculation for cost-effectiveness with the attributes health improvement and drug costs. Since drugs could have high costs and large health improvements, and still be on the edge

of cost-effectiveness, the author chose to avoid this problem and included the cost attribute as cost-effectiveness. The DCE was designed in such a way that the health improvement indicates the size of the effect of the drug and the cost-effectiveness indicates if the price is good or bad, relative to the stated effectiveness. The budget impact is not only linked to the price, but also to the epidemiology of the disease, which was not the purpose of this attribute. As mentioned before, numerical levels were not included to prevent a high cognitive burden. A bad, moderate and good cost-effectiveness represented the different levels.

*Composition of health gain*

This attribute showed what health improvement entails. A drug can improve QoL, provide EoL or both, which represented the different levels for this attribute. The composition of health gain is an attribute that is not currently used in the HTA process in the Netherlands and also not widely used in the literature, although it was used by Koopmanschap et al., also in a Dutch health care setting (13,21). The current reimbursement process looks at the number of QALYs. EoL without a reduction in the QoL results in more QALYs, which implicitly includes the EoL in the effectiveness evaluation. This could also be done explicitly, by weighting QoL and EoL. If society places more weight on either the EoL or the QoL, the composition of health gain could be a valuable implementation in the reimbursement process of new drugs.

*Age of the patient population*

Age was included in a DCE held in members of the NICE and was ranked the least important attribute by HTA decision makers (22). Health care costs are expected to rise in the coming decades and one of the factors associated with rising health care expenditure is aging (2,28). Students might reflect this in a preference for less expensive treatments targeted to people of a certain age, which makes this attribute an interesting inclusion. The attribute age ranked fifth important in the conducted pilot study. The DCE included the age groups children, adults and elderly people.

Age is currently not an explicit criterion in the reimbursement process, but it is incorporated in the proportional shortfall calculation. If for example, a child and an elderly person both have a disease that results in the loss of 2 remaining life years, the proportional shortfall of the elderly person will be higher. This is because the proportional shortfall calculates the relative loss in QALYs and the elderly person only had a few years remaining without the disease. The elderly person would have the highest disease severity in this case. It is however stated by ZIN that recently, people seem to prefer health gains in younger people over older people, which argues against the current calculation method of the proportional shortfall (29).

*Size of the patient population*

The final included attribute, population size, was included in the DCE constructed by Wranik et al. and contained the levels a high and low number of patients. Wranik et al. found no significant effect on population size alone, although an interaction effect showed that people were willing to accept a non-superior survival benefit when the population size was large (23). The attribute ranked fourth important in the conducted pilot, showing an indication that people placed some value on this attribute. The reason and the direction of this interest were unfortunately not captured in the pilot. People could argue against a larger population, thinking about the costs that might be associated with larger populations. On the other hand, larger populations also result in larger total health gains. The goal of this attribute was to elicit whether people placed any importance on the size of the population and which direction this preference would be.

Like the other attributes, the level selection for the attribute size of the patient population was kept simple. The selected levels were a small, moderate and large patient population.

*Table 2. Final selection of attributes and levels.*

| **Attributes** | **Levels** |
|---|---|
| Disease severity | Low impediment<br>Moderate impediment<br>High impediment |
| Health improvement | Small<br>Moderate<br>Large |
| Composition of health gain | 50% EoL and 50% QoL<br>100% EoL<br>100% QoL |
| Cost-effectiveness | Bad<br>Moderate<br>Good |
| Size of the patient population | Small<br>Moderate<br>Large |
| Age of the patient population | Children (0 – 18 years)<br>Adults (18 – 67 years)<br>Elderly ( > 67 years) |

*QoL Quality of life, EoL extension of life.*

## 3.2 Experimental design

The experimental design determines how much statistical information can be obtained from the DCE. Orthogonal designs are efficient when the underlying statistical model assumes linearity. In the case of a nonlinear statistical model, like the Conditional logit model, a D-efficient design is more efficient. D-efficient designs seek to minimize average parameter-estimate variances. As for the number of choice tasks a single person can complete, 8 to 16 choice tasks are seen as good practice (27).

Two fractional D-efficient designs, one with naïve priors and one with informed priors, of 24 choice tasks divided into two blocks were constructed with the software programme Ngene. Each choice task had two alternatives and no opt-out option. It was expected that few would choose the opt-out because there was no personal downside for choosing the opt-out. The opt-out would therefore have provided little information. The design with naïve priors consisted out of uniform priors based on identified literature and the conducted qualitative pilot. To improve the efficiency of the design, the design was updated after roughly 30 to 40 respondents completed the first design (30). The informed priors of the second version of the design were specified Bayesian normally distributed priors, based on the results of a mixed logit (MIXL) model, for the attributes that already showed significant results. The priors of insignificant coefficients were updated using the 95% confidence intervals with an uniform distribution if the results were in line with expectations. When the results deviated from the expectations, the updated priors would be a combination of the results and the initial prior based on the literature and conducted pilot. The design included no interactions because the number of parameters was already substantial and interactions would increase this number even more, which could lead to too little statistical power. All possible scenarios were seen as plausible by the author, therefore no scenarios were prohibited in the design generation. The Ngene syntax and priors of the designs are listed in appendix 3.

To ensure enough statistical power the sample size also needs to be large enough. The optimal sample size is dependent on multiple factors, for example, the level of certainty, the magnitude of the expected differences, the purpose of the research and the methods that will be used. Sample sizes generally range from roughly 150 to 1,200 respondents (31). Larger sample sizes of 200 respondents per group are needed when multiple groups are compared with the goal of detecting significant differences. 300 respondents are recommended for robust quantitative research. For investigational research and developing new hypotheses only 30 - 60 respondents are needed (31).

Johnson and Orme proposed a rule of thumb to determine minimum sample sizes for aggregate level full-profile choice-based conjoint analyses. The proposed rule of thumb is listed in the equation below where N equals the number of respondents, t equals the number of choice tasks, a equals the number of alternatives for each task and c is equal to the largest number of levels for any attribute (31). When applied to the design and characteristics of this DCE (equation *III*), a sample size of at least 63 respondents was recommended.

$$N > \frac{500c}{t*a} \qquad N > \frac{500*3}{12*2} \qquad III$$

The sample size is also dependent on the available resources. Larger sample sizes tend to be harder to achieve in practice (31). This research had a tight timeline and used convenience sampling to gather

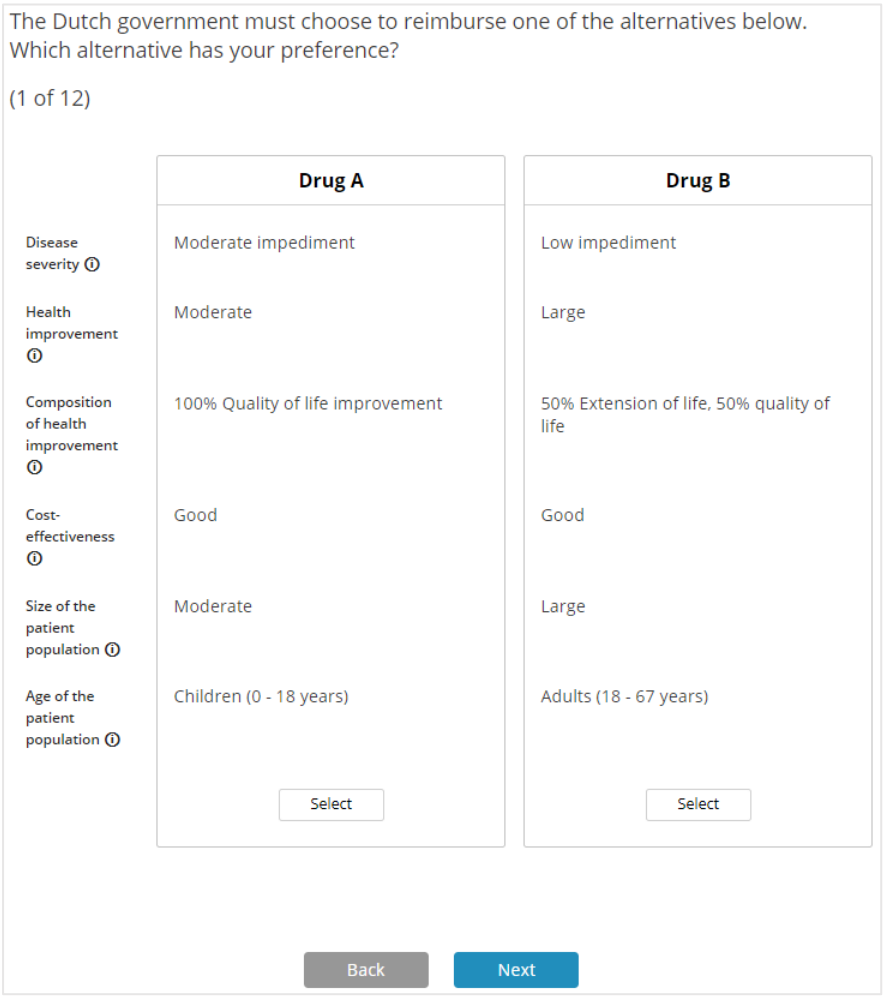
respondents. With that in mind, large sample sizes might not be practically achievable. Therefore, the minimum sample size is set on 100 respondents, which is substantially higher than the rule of thumb suggested, also aiming for the lower limit of the general range of DCEs, while still a practically achievable number considering the before mentioned constraints.

## 3.3 Survey design

The survey was created with the programme Sawtooth and was written in Dutch. It was expected that most people in the sample population would be Dutch and that a Dutch survey would improve the willingness to complete the survey and the understandability of the survey. The survey was divided into multiple blocks, starting with an introduction explaining the goal of the survey and subsequentially containing an anonymity statement. The next part of the survey included a table of contents explaining the survey structure, followed by questions about age and gender. After that, two pages explaining the different attributes and levels followed, where each page contained an example choice task. These example choice tasks both had a dominant option and were used as a consistency check whether respondents chose the dominant option. Within-set dominant pairs is a method to test the internal validity where one of the alternatives contains better or dominant levels across all attributes compared to the lesser alternative (32). The validity test would lower the efficiency of the design when it would be placed in the actual choice task and is therefore placed in the example choice tasks.

The 12 choice tasks of the DCE started after the explanation of the attributes. Each attribute in each choice task had a button where respondents could click or hover their mouse to see a pop-up containing information about the corresponding attribute. An example DCE choice task is shown in figure 3. After the 6$^{th}$ choice task, the respondents were noticed that they were halfway through the DCE and had to answer two questions regarding their education level and education background. The survey ended with 4 evaluation questions and an open question to leave comments about the survey. The four evaluation questions evaluated the understandability of the survey and used a Likert scale where respondents had to choose between 1 to 5 where 1 was equal to strongly disagree and 5 was equal to strongly agree.

*Figure 3. Example choice task of the DCE, translated to English*




## 3.4 Data collection

Data were obtained by online distribution of the survey using Sawtooth's hosting function. The sample is a convenience sample of students recruited using the authors' network of friends and fellow students, and by distributing the link to the survey at the Utrecht University campus. First, respondents were recruited to complete the first design. After the design was updated, new respondents were recruited to complete the updated design. Students at vocational education (MBO), applied sciences (HBO) and research university (WO) were included if they spoke well enough Dutch to understand the survey. Exclusion criteria were people who did not study and students with an insufficient level of the Dutch language to understand the survey.

## 3.5 Statistical analysis

The results were analysed with the programme StataMP 16 and p-values $< 0.05$ were considered as significant. A table with descriptive characteristics of all participants was made, including the results of the cognitive evaluation questions. Answering times were assessed to identify speeders and respondents were classified as speeder when they finished the survey below 0.5 * median time of survey

completion. P-values for differences in demographic variables between the respondents who completed the first and second design were calculated with independent group t-tests for continuous variables and with Chi-squared tests or Pearson's chi-squared test for binary variables. A Pearson's chi-squared test was used when a number < 5 was present in the contingency tables. Differences between the demographic characteristics of the respondents that completed the two different designs were assessed because the second design was more efficient. Changes in the demographic characteristics between the two groups that filled in the different designs might have an influence on the results.

The choice task results were analysed with a Latent class (LC) model and a MIXL model. The equation for the models is shown in equation *IV*. The distinction between the models is the method for dealing with heterogeneity. A LC model constructs multiple classes and assumes that these classes are sufficient to capture the differences in preferences within the population. A MIXL model assumes that people have different preferences and captures the preference heterogeneity by estimating standard deviations (SD) for the preferences (33) and was performed with 500 Halton draws and the bfgs technique. The selection of the preferred model (LC vs MIXL) was based on the model fit and how well the model was capable of answering the research questions.

$$V_{altA} = V_{altB} = \beta_{1-3} * \text{disease severity} + \beta_{4-6} * \text{health improvement} + \beta_{7-9} * \text{composition of health gain} + \beta_{10-12} * \text{cost} - \text{effectiveness} + \beta_{13-15} * \text{population size} + \beta_{16-18} * \text{age of population} \qquad IV$$

The attributes and levels were included as categorical variables and were dummy coded for the analysis. A separate variable was constructed for each attribute level, where the values would be 0 or 1, depending on if that attribute level would be in the particular choice task. For example, disease severity was divided into three separate variables: low disease severity, moderate disease severity and high disease severity. The final model included only two of these variables, because the third variable (in this case low disease severity) was the reference level. The beta coefficients then will provide the attribute preferences and therefore provide information for answering research question 1.

Interaction terms for students with a medical background were added to the MIXL model. This will provide information on preference differences between students with a medical background and students with other backgrounds, which will be used for answering research question 2. A model containing interaction terms for all variables might contain too many parameters to construct. A selection of the most important interactions was when including all interactions to the model was not feasible. In this case, the most important interactions will be selected based on different models containing only a few interaction terms to make the model feasible. To evaluate the effect of the speeders, a sensitivity analysis excluding speeders was performed with the preferred model.

Choice predictions and marginal effects were calculated with the results of the selected model to enhance the interpretability of the results. The formula used for the calculation of the choice predictions

is shown in equation *V*. Five choice predictions were calculated for the following scenarios compared to a base case that consisted of moderate or standard levels for all attributes in adults:

- a best scenario, using the levels with the highest expected utility;
- a worst scenario, using the levels with the lowest utility;
- a low cost-effective drug with a good health improvement for sick elderly people;
- a highly cost-effective drug with moderate health improvement for moderately sick children;
- a moderate cost-effective drug with moderate health improvement for a large population of sick adults.

$$\text{Probability} = \frac{\exp(U_1)}{\exp(U_1) + \exp(U_{base\ case})} \qquad V$$

# Chapter 4. Results

## 4.1 Collected data and sample characteristics

In total, 115 respondents completed the survey from April 2021 to June 2021, of which 84 respondents completed the first design and 31 respondents completed the updated design. The design was updated later than initially planned due to a hosting problem of the survey. The demographic information of the respondents is listed in table 3. The mean age was 22.9 years old. Respondents were predominantly female (59.1%) and most people were currently educated at a master's university level (60.9%). 59.1% of the respondents were following education with a medical background.

The educational background differed significantly between the respondents who completed the first and second design of the DCE. The respondents that filled in the first design mostly enjoyed a medical background (59.1%), followed by a technical (13.9%) or social background (13.0%). The group that filled in the second design had a lower proportion of students with a medical background (35.5%), although it was still the most prevalent educational background. Technical (29.0%), social (16.1%) and other backgrounds (16.1%) were most present after the medical background (35.5%) in the respondents that completed the second design.

92.2% of all respondents correctly completed the two dominant example scenarios in the explanation part of the survey. 7.0% correctly answered one of the two scenarios and 1 respondent failed to pick any of the dominant scenarios. The mean and median time respondents used for the completion of the survey were 20.8 and 10.5 minutes, respectively. 6.1% of the respondents were classified as speeders and completed the survey within half the median time.

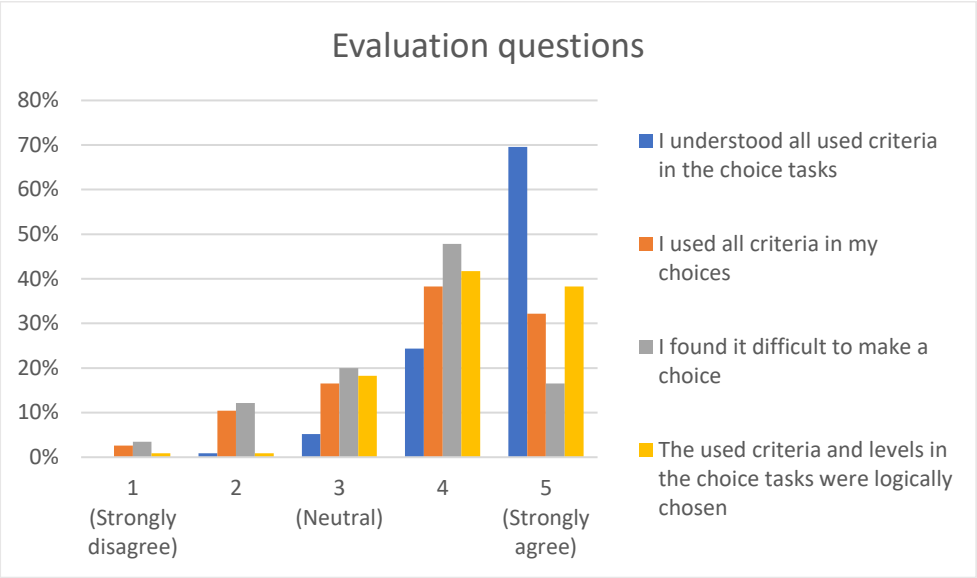
*Table 3. Demographic information of the respondents. Speeders were defined as respondents who completed the survey below 0.5 * median completion time.*

| | | Total | (%) | First design | (%) | Second design | (%) | P- value |
|---|---|---|---|---|---|---|---|---|
| Total | | 115 | 100.0% | 84 | 73.0% | 31 | 27.0% | |
| Age (mean) (SD) | | 22.9 (2.1) | | 22.8 (2.1) | | 23.1 (2.2) | | 0.506[1] |
| Male | | 47 | 40.9% | 37 | 44.0% | 10 | 32.3% | 0.254[2] |
| Female | | 68 | 59.1% | 47 | 56.0% | 21 | 67.7% | |
| Education level | University master | 70 | 60.9% | 51 | 60.7% | 19 | 61.3% | 0.166[3] |
| | University bachelor | 25 | 21.7% | 20 | 23.8% | 5 | 16.1% | |
| | HBO | 18 | 15.7% | 13 | 15.5% | 5 | 16.1% | |
| | MBO | 2 | 1.7% | 0 | 0.0% | 2 | 6.5% | |
| Education background | Medical | 68 | 59.1% | 57 | 67.9% | 11 | 35.5% | 0.002[3] |
| | Economical | 7 | 6.1% | 6 | 7.1% | 1 | 3.2% | |
| | Technical | 16 | 13.9% | 7 | 8.3% | 9 | 29.0% | |
| | Social | 15 | 13.0% | 10 | 11.9% | 5 | 16.1% | |
| | Cultural | 1 | 0.9% | 1 | 1.2% | 0 | 0.0% | |
| | Other | 8 | 7.0% | 3 | 3.6% | 5 | 16.1% | |
| Dominant questions | Both correct | 106 | 92.2% | 77 | 91.7% | 29 | 93.5% | 0.544[3] |
| | One correct | 8 | 7.0% | 6 | 7.1% | 2 | 6.5% | 0.631[3] |
| | None correct | 1 | 0.9% | 1 | 1.2% | 0 | 0.0% | 0.730[3] |
| Speeders | | 7 | 6.1% | 4 | 4.8% | 3 | 9.7% | 0.385[3] |

*SD standard deviation, HBO applied sciences, MBO vocational education, [1] student's t-test, [2] Chi-square test, [3] fishers exact test.*

The results of the evaluation questions are presented in figure 4. The understandability of the used criteria in the choice tasks was high. 93.9% of the respondents ranked the understandability 4 and 5 out of 5. 70.5% scored 4 or 5 on the question whether they used all criteria in their choices and 64.3% scored 4 or 5 whether they found it difficult to make a choice. The levels and choice tasks were logically chosen according to 80.0% of the respondents who ranked this 4 or 5 out of 5.

*Figure 4. Bar graph showing the results of the evaluation questions*




## 4.2 Model selection

The data were analysed with a MIXL and a LC model. Goodness of fit of both models is presented in table 4. The LC model was performed with 2 classes since the number of respondents for each group would be very small when dividing the obtained sample into 3 classes. Although the LC model showed a better AIC and BIC compared with the MIXL model, the latter was selected for the main model, motivated by the following: the LC model showed some differences in demographic factors between the classes, but not enough to explain the differences between the classes well. The MIXL model also connected well to the research questions because a MIXL model is more straightforwardly able to examine the differences in preferences between medical and non-medical students with the inclusion of interaction terms to the model. The results of the LC model are presented and discussed shortly in appendix 4.

*Table 4. Model fit of the Mixed logit model and the Latent class model.*

| Model | AIC | BIC |
|---|---|---|
| Mixed logit model | 1,448.386 | 1,590.538 |
| Latent class model (2 classes) | 1,430.1578 | 1,498.7811 |
| Latent class model (3 classes) | 1,418.0505 | 1,522.3579 |

*AIC = Akaike information criterion, BIC = Bayesian information criterion.*

## 4.3 Results of preferred mixed logit model

The results of the MIXL model are presented in table 5. The ranges of the attribute preferences with the corresponding standard deviations are graphically shown in figure 5. All attributes were significant ($p < 0.01$), except for the composition of health gain, which did not have a significant coefficient for both levels. Although the coefficients were not significant, the level '100% EoL' showed a significant amount of heterogeneity, indicating that there were differences in preferences present among the

students. Students ranked a treatment aimed at elderly people as the most important attribute with a coefficient of -1.96, indicating a preference for treatments aimed at younger people. This level also showed the highest heterogeneity, SD 1.10, of all levels across the attributes. The most important attributes after the treatment age attribute were cost-effectiveness, health improvement and disease severity. Within these attributes, students preferred more cost-effective treatments, large health improvements and treatments aimed at people with a high disease severity. Students also showed an increasing preference for larger population sizes.

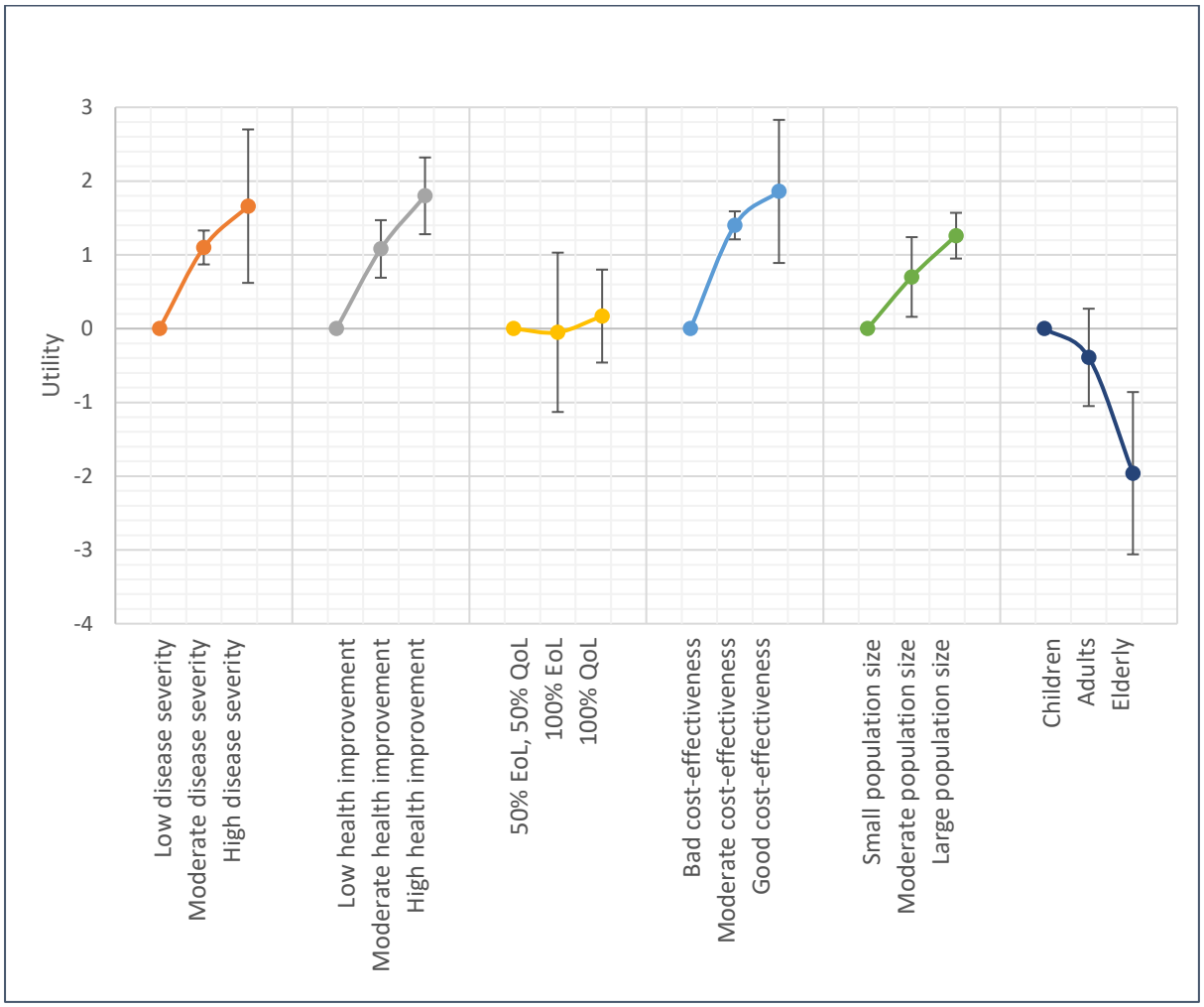
The heterogeneity for cost-effectiveness, health improvement and disease severity were significant ($p < 0.05$) for the upper-end levels of these attributes. The distribution of the preference of good cost-effectiveness (figure 6) showed that the preference for the level good cost-effectiveness had a normal distribution, which was slightly positively skewed. Other levels with a significant heterogeneity were 100% QoL, 100% EoL and an adult target population. A sensitivity analysis excluding speeders showed similar results, except for one significant change in the preference heterogeneity for a treatment aimed at adults. The results of this sensitivity analysis are shown in appendix 5.

*Table 5. Results of the main mixed logit model. standard errors in parentheses.*

| Attribute | Attribute levels | Coefficient (SE) | SD (SE) |
|---|---|---|---|
| Disease severity | Low disease severity | Reference | |
| | Moderate disease severity | 1.095[1] (0.165) | 0.225 (0.326) |
| | High disease severity | 1.655[1] (0.233) | 1.040[1] (0.210) |
| Size of the health improvement | Small health improvement | Reference | |
| | Moderate health improvement | 1.075[1] (0.165) | -0.391[3] (0.223) |
| | Large health improvement | 1.804[1] (0.238) | 0.518[2] (0.237) |
| Composition of health gain | 50% EoL, 50% QoL | Reference | |
| | 100% EoL | -0.0466 (0.173) | 1.082[1] (0.216) |
| | 100% QoL | 0.166 (0.143) | 0.627[1] (0.205) |
| Cost-effectiveness | Bad cost-effectiveness | Reference | |
| | Moderate cost-effectiveness | 1.404[1] (0.186) | -0.186 (0.231) |
| | Good cost-effectiveness | 1.859[1] (0.246) | 0.966[1] (0.207) |
| Population size | Small population size | Reference | |
| | Moderate population size | 0.697[1] (0.153) | -0.536[2] (0.209) |
| | Large population size | 1.264[1] (0.171) | -0.306 (0.272) |
| Age of the treatment population | Children | Reference | |
| | Adults | -0.392[1] (0.143) | 0.661[1] (0.215) |
| | Elderly | -1.963[1] (0.246) | 1.104[1] (0.240) |

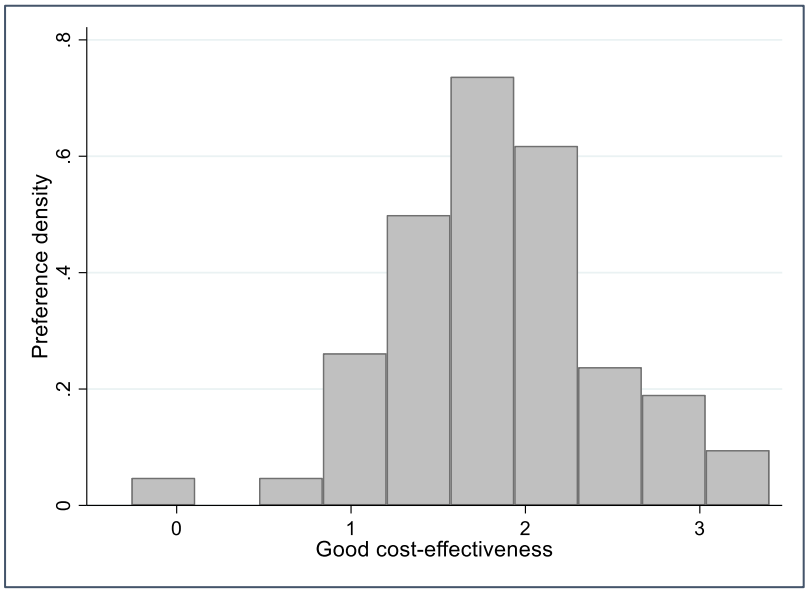
*[1] p<0.01, [2] p<0.05, [3] p<0.1, SD standard deviation, SE standard error, QoL quality of life, EoL extension of life.*

*Figure 5. Range of the attribute preferences including the reference levels (± standard deviation).*




*EoL Extension of life, QoL quality of life.*

*Figure 4. Histogram of the preference distribution for good cost-effectiveness.*



## 4.4 Choice predictions and marginal effects of the mixed logit model

Choice predictions were calculated to find the likelihood of people preferring certain scenarios and are listed in table 6. A best and worst scenario were calculated and the results showed that these scenarios would be chosen 94.6% and 0.3%, respectively, when compared to the presented base case scenario. The third scenario showed that although preferences for disease severity and health improvement were very strong, the choice probability was still only 42.8% because the treatment was aimed at elderly people. When a treatment was cost-effective and aimed at children, the probability of being chosen increased to 73.5%. Also, population size and disease severity can be drivers of improved acceptability, which was seen in the choice probability of scenario 5 (75.4%).

*Table 6. Choice probabilities for five scenarios compared to the base case.*

| | Base case | Scenario 1 (best) | Scenario 2(worst) | Scenario 3 | Scenario 4 | Scenario 5 |
|---|---|---|---|---|---|---|
| Choice probability (%) | - | 94.6% | 0.3% | 42.8% | 73.5% | 75.4% |
| Disease severity | Moderate | High | Low | High | Moderate | High |
| Health improvement | Moderate | High | Low | High | Moderate | Moderate |
| Composition of health improvement | 50% EoL, 50% QoL | 100% QoL | 100% length | 50% EoL, 50% QoL | 100% QoL | 50% EoL, 50% QoL |
| Cost-effectiveness | Moderate | High | Low | Moderate | High | Moderate |
| Population size | Moderate | Large | Small | Moderate | Moderate | high |
| Age | Adults | Children | Elderly | Elderly | Children | adults |

*EoL extension of life, QoL quality of life.*

Marginal effects were calculated for single-level changes in scenario 1 (best case) compared to the base case and are shown in table 7. The scenario change from children to elderly, a good to a bad cost-effectiveness, a large to a small health improvement and a high to a low disease severity all had a similar impact ranging from a marginal effect of -17.7% to -23.5%. The change from a large to a small population size decreased the choice probability by -11.4% and a change in the composition of health gain only had an impact of -1.2% on the choice probability.

*Table 7. Marginal effects on the choice probability for scenario 1 (best case) compared with the base case after changing a single attribute level.*

| Level change | Marginal effects |
|---|---|
| High to low disease severity | -17.7% |
| Large to small health improvement | -20.3% |
| 100% QoL to 100% EoL | -1.2% |
| Good to bad cost-effectiveness | -21.5% |
| Large to small population size | -11.4% |
| Children to elderly target population | -23.5% |

*QoL quality of life, EoL extension of life.*

## 4.5 Differences between students with a medical or other educational backgrounds

A MIXL model containing all variables and interaction terms for the variables for students with a medical background could not be constructed because it exceeded the maximum amount of 20 independent variables for a MIXL model. Three separate MIXL models were constructed that included only four of the interaction term variables. The interactions with the lowest impact, based on the p-values, were excluded from the final model to limit the number of included variables to a maximum of 20. Attribute level interactions excluded from the final model were moderate disease severity, moderate health improvement, 100% EoL and 100% QoL. The final model with all included subgroup interactions is shown in table 8.

None of the included interaction terms was significant with the prespecified p-value < 0.05, although the subgroup interaction with a good cost-effectiveness and an elderly population had p-values < 0.1. These interactions showed that, although not significant, medical students seemed to prefer treatments with a good cost-effectiveness (coefficient 2.181) more compared to students with another educational background (coefficient 1.528). The heterogeneity for students with a medical background was also significant for treatments with a good cost-effectiveness. Medical background students showed lower negative preferences for the elderly population compared to students with other backgrounds, coefficients -1.758 and -2.406, respectively.

*Table 8. Results of the mixed logit model with the most important interactions for medical study background. Standard errors in parentheses.*

| Attributes | Attribute levels | Mean (SE) | SD (SE) |
|---|---|---|---|
| Disease severity | Low disease severity | Reference | |
| | Moderate disease severity | 1.130[1] (0.170) | -0.313 (0.292) |
| | High disease severity | 1.603[1] (0.250) | 0.216 (0.349) |
| | Medical * high disease severity | 0.249 (0.343) | 1.474[1] (0.326) |
| Size of the health improvement | Low health improvement | Reference | |
| | Moderate health improvement | 1.084[1] (0.180) | 0.412[3] (0.226) |
| | Large health improvement | 1.601[1] (0.265) | -0.575 (0.425) |
| | Medical * large health improvement | 0.453 (0.298) | -0.314 (0.666) |
| Composition of health gain | 50% EoL, 50% QoL | Reference | |
| | 100% EoL | -0.0114 (0.181) | 1.124[1] (0.233) |
| | 100% QoL | 0.212 (0.144) | -0.483[3] (0.288) |
| Cost-effectiveness | Bad cost-effectiveness | Reference | |
| | Moderate cost-effectiveness | 1.197[1] (0.235) | -0.0903 (0.246) |
| | Medical * moderate cost-effectiveness | 0.486 (0.344) | -0.334 (0.546) |
| | Good cost-effectiveness | 1.528[1] (0.285) | 0.575[3] (0.319) |
| | Medical * good cost-effectiveness | 0.653[3] (0.381) | 1.256[1] (0.459) |
| Population size | Small population size | Reference | |
| | Moderate population size | 0.479[2] (0.209) | 0.557[2] (0.217) |
| | Medical * moderate population size | 0.389 (0.282) | 0.104 (0.382) |
| | Large population size | 1.157[1] (0.212) | -0.0947 (0.318) |
| | Medical * large population size | 0.240 (0.285) | -0.309 (0.384) |
| Age of the treatment population | Children | Reference | |
| | Adults | 0.548[1] (0.189) | 0.322 (0.340) |
| | Medical * adults | 0.163 (0.299) | 1.009[1] (0.340) |
| | Elderly | 2.406[1] (0.329) | 1.004[1] (0.213) |
| | Medical * elderly | 0.648* (0.385) | 0.205 (0.383) |

*[1] p<0.01, [2] p<0.05, [3] p<0.1, SD standard deviation, SE standard error, QoL quality of life, EoL extension of life.*

# Chapter 5. Discussion and conclusion

## 5.1 Summary and context of main findings

This study attempted to estimate preferences of Dutch students for different criteria which are potentially used in the HTA process for drugs. Currently used criteria were evaluated, but also criteria that do not have a place in the current HTA process. Research like this is needed when updating the HTA frameworks, to ensure that the HTA process still reflects the preferences of society. In this study, students ranked age, cost-effectiveness, the size of health improvement and the disease severity as the most important attributes. The attribute population size was ranked less important, although still significant. The composition of health gain was the only attribute that did not have a significant preference. The currently used criteria in the Dutch reimbursement process all ranked fairly high, which confirms that the current HTA process is at least partly supported by students. This is in line with earlier work in the general public, who valued similar attributes, but also harmonious with preferences of Dutch HTA stakeholders, who valued disease severity, cost per QALY gained, individual health gain and budget impact as the most important attributes (21,24). Caution is needed when comparing results because other DCEs had different attributes and level selections which influenced the trade-offs. The attributes disease severity and cost-effectiveness showed a flattening concaveness in the incremental effects from the lowest to the highest level. This indicates that people did not desired a drug with a bad cost-effectiveness or a drug for people with a mild disease severity, but that their incremental preferences decreased when the change was for a treatment with a moderate to a good cost-effectiveness or moderate to a severe disease severity.

The students did however ranked the attribute age, which is not explicitly included in the current reimbursement process, as the most important attribute. Students had a clear preference for the reference category children, a small negative preference for adults and a large negative preference for elderly people. This was contradicting with earlier findings in NICE committee members, where age did not have a significant effect (22). Our finding that students preferred treatments aimed at younger people is not in line with the proportional shortfall calculation for disease severity currently used in the reimbursement process, which implicitly allocates a higher weight on older people. Reckers-Droog et al. already showed societal concerns that age might not be correctly reflected in the proportional shortfall calculation. An adjustment for age could be incorporated to reflect the societal preference for different age groups, although more empirical evidence on age preferences was needed to check whether such an adjustment would better align with public preferences (29).

A reason for the large age preference that was found could relate to the sample population. Respondents were only 22.9 years old on average, which implies that elderly care will not take place for themselves in the near future. Students might discount effects in the future and therefore place lower importance on treatments taking place in their distant future. However, this effect was not seen in the levels QoL

and EoL of the attribute composition of health gain, since this attribute was not significant. This was contradicting with the results of Koopmanschap et al. who found a significant result for the attribute composition of health gain in HTA shareholders and HTA students. In particular, HTA students valued QoL over EoL in the work of Koopmanschap et al (21).

Another attribute that showed significance was the size of the patient population, although it was ranked less important compared to the previously mentioned attributes. In general, students preferred larger population sizes. The reason for this could be the larger total health improvement that is gained when treating a larger population. This significant result was contradicting to the earlier work in HTA experts by Wranik et al., who did not found a significant effect. This difference might be partly explained by the differences in the sample population. Wranik et al. questioned HTA experts who might associate a larger population not only with more health gains, but also with an increase in costs and therefore not use a large population as a reason to reimburse a drug. Our study was held in students, who might think less about the costs involved and more about the obvious larger health gains associated with a larger population. The differentiation between medical students and students with another educational background had no significant impact on the preference for population size.

The finding on population size is in line with earlier work that estimated preferences of the general public, although these results were not perfectly concise (34). However, this DCE was focused at orphan drugs, where population sizes play a more important role. Societal preferences consistently shown to not prefer extra funding allocation towards rare diseases based on rarity alone when funding decisions were traded against treatments for common diseases (35). The found preference for population size in our work shows a similar trend, preferring larger population sizes and pursuing the highest total health gain.

In contrast to a Conditional logit model, a MIXL model provides information on the heterogeneity of the preferences in the sample. There was a significant amount of heterogeneity in the attribute levels high disease severity, 100% QoL improvement, 100% length of life improvement, high cost-effectiveness, an adult population and an elderly population. This heterogeneity indicated that there were differences in the preferences in the sample for these attribute levels. This is also interesting with the concaveness earlier discussed for the attributes disease severity and cost-effectiveness. The sample had different preferences for these decreasing incremental preferences. The distribution of the preference for high cost-effectiveness was slightly positively skewed, which indicated that there were people in the sample who did not share the preference for this decreasing incremental effect on cost-effectiveness. The distribution of the disease severity does not have a similar pattern.

No significant differences were found between students with a medical and non-medical background after adding interactions to the MIXL model. The interactions for an elderly population and a treatment with a good cost-effectiveness were the only interactions close to significance ($p < 0.1$). Medical

students showed a non-significant, higher interest in drugs with a good cost-effectiveness and a lower negative interest in elderly populations compared to students with other backgrounds. The direction of these differences is consistent with the results of Tappenden et al., where age was not a significant attribute for HTA shareholders (22). Medical students might have more background knowledge on the HTA process and might therefore place more emphasis on current HTA criteria, such as cost-effectiveness, and less on equity factors, such as age.

Furthermore, the internal validity of the study was tested with two within-set dominant choice tasks in the example choice tasks. Because the test was placed in the explanation part of the survey, people might be less focused compared to when completing the actual choice tasks, which might result in a less accurate validity test. The dominant option was also not fully dominant because the attribute composition of health gain did not have a dominant option. This was expected to only have a minor influence, since the choice tasks were fully dominated by the other attributes. This expectation was confirmed by the results of the MIXL model, which showed that the composition of health gain was non-significant, and by the results of the validity test, which showed a good validity and only one respondent failed to answer both dominant questions correctly. In addition to the validity test, debriefing questions were asked to check whether the respondents understood the DCE. The questions showed that the DCE was generally well understood and that most people considered all attributes when choosing an alternative.

## 5.2 Limitations and strengths of the analysis

This study has several limitations. First, the experimental design was not updated until 84 respondents had completed the survey. Initially, the design was supposed to be updated after roughly 30 respondents to achieve a higher efficiency with a more efficient design. A problem concerning the hosting process of the survey resulted in the delayed implementation of the updated design. The completion of the first design by more respondents most likely resulted in a more efficient updated design, because the coefficients were estimated more accurately. Unfortunately, this updated design was only completed by 31 respondents, which negatively influenced the power of the study. Too efficient designs can have downsides as well, for example, respondents not trading across the attributes, but using simplistic heuristics to choose their preferred scenario (36).

Another limitation is the small sample size of the population, which in combination with the lowered efficiency might have caused too little power to show statistical differences between medical students and students with other backgrounds. Larger sample sizes of 200 respondents for each group are preferred when the goal is to detect significant differences in preferences between multiple groups (31). Third, the external validity was not assessed during this study. This is a limitation in the majority of performed DCEs in the area of health economics and is therefore an opportunity for improvement (19). External validity tests are difficult to implement for this subject in the general public since they have

not made reimbursement decisions in real life. A DCE held among ZIN employees could use ZIN reimbursement advices as external validity testing, comparing stated preferences with revealed preferences. Finally, because the sample was not experienced in the subject, general and understandable attributes had to be used. This resulted in the exclusion of a numerical price variable, which made it impossible to calculate a willingness to pay.

A strength of this study is the used methodology. Compared to earlier work performed in the same institutional setting this study used a more sophisticated model which provided additional information on the heterogeneity (21). Another strength of this study is the composition of the sample. The sample consists for 59.1% of medical students and 40.9% of students with a different background. This gives the study a unique composition of people that might become health care professionals in the near future and people that will only experience the health care sector as patients or general public. The health care sector should ideally embody a combination of preferences of the general public and health care professionals. Public input is desirable and should have a meaningful role in healthcare prioritizing (37). Differences between students' preferences with and without a medical background are comparable in this study and those differences are valuable insights for policymakers when constructing drug reimbursement guidelines. Future research can build on this study and explore the differences between health care professionals and the general public in more detail. Preferences of other subgroups of the population must be elicited and compared to the results of this study to acquire a complete view of the preferences of the general public.

## 5.3 Conclusion

This study found that students ranked both currently and not currently used criteria important for the reimbursement process of drugs. The most important criterion was the age of the treatment population, which is a criterion not explicitly used in the current Dutch reimbursement process. An especially strong negative preference was seen for elderly people. If future research conducted in different age groups confirms this finding, adding age adjustments to the proportional shortfall calculation could be a solution to improve the alignment of the reimbursement process with public preferences. Students showed similar significant preferences for disease severity, size of the health improvement and cost-effectiveness, which are all criteria widely used in the HTA process. This indicates that the current Dutch reimbursement process already reflects a large part of the preferences of students. Additionally, students showed a significant preference for larger populations, but not for the composition of health gain. No differences were seen between the preferences for medical students compared to other students, although this was perhaps due to a sample size that was too low. DCEs are an excellent tool for eliciting preferences in the health care system and provide valuable information for the development of HTA guidelines. This study adds new information in this area and provides exploratory insights in differences between students with medical and other educational backgrounds, showing potential areas

of preference conflicts between health care professionals and patients, which must be taken into account when constructing HTA guidelines.

# References


1. RIVM. Zorguitgaven blijven tot 2060 stijgen, gemiddeld met 2,8 procent per jaar | RIVM [Internet]. [cited 2021 May 25]. Available from: https://www.rivm.nl/nieuws/zorguitgaven-blijven-tot-2060-stijgen-gemiddeld-met-28-procent-per-jaar

2. Dieleman JL, Squires E, Bui AL, Campbell M, Chapin A, Hamavid H, et al. Factors associated with increases in US health care spending, 1996-2013. JAMA - J Am Med Assoc. 2017 Nov 7;318(17):1668–78.

3. Zorginstituut Nederland. GIPdatabank [Internet]. [cited 2021 Feb 12]. Available from: https://www.gipdatabank.nl/

4. Hirsch BR, Balu S, Schulman KA. The Impact Of Specialty Pharmaceuticals As Drivers Of Health Care Costs. Health Aff. 2014;33(10):1714–20.

5. Sorenson C, Drummond M, Kanavos P. Ensuring value for money in health care The role of health technology assessment in the European Union. WHO regional Office Europe; 2008.

6. Viney R, Lancsar E, Louviere J. Discrete choice experiments to measure consumer preferences for health and healthcare. Expert Rev Pharmacoecon Outcomes Res. 2014;2(4):319.

7. Jakubiak-Lasocka J, Eng M, Jakubczyk M. Cost-effectiveness versus Cost-Utility Analyses: What Are the Motives Behind Using Each and How Do Their Results Differ?-A Polish Example. Value Heal Reg Issues. 2014;4:66–74.

8. Schmitz S, Mccullagh L, Adams R, Barry • Michael, Walsh C. Identifying and Revealing the Importance of Decision-Making Criteria for Health Technology Assessment: A Retrospective Analysis of Reimbursement Recommendations in Ireland. Pharmacoeconomics. 34.

9. Akehurst RL, Abadie E, Renaudin N, Sarkozy F. Variation in Health Technology Assessment and Reimbursement Processes in Europe. Value Heal. 2017 Jan 1;20(1):67–76.

10. Bouvy JC, Cowie · Luke, Lovett · Rosemary, Morrison D, Livingstone H, Crabb · Nick. Use of Patient Preference Studies in HTA Decision Making: A NICE Perspective. Patient - Patient-Centered Outcomes Res. 2020;13:145–9.

11. Juhnke C. Patient Preferences Versus Physicians' Judgement: Does it Make a Difference in Healthcare Decision Making? Appl Health Econ Health Policy. 2013;11(3):163–80.

12. Lancsar E, Louviere J. Conducting Discrete Choice Experiments to Inform Healthcare Decision Making A User's Guide. Vol. 26, Pharmacoeconomics. 2008.

13. Vijgen S, Heesch F van, Obradovic M. Ziektelast in de praktijk. Dutch Natl Heal Care Inst.

2018;1–34.

14. VWS. Assessment procedure of outpatient medicines [Internet]. 2018. Available from: www.zorginstituutnederland.nl

15. Dutch Institute National Health Care (Zorginstituut Nederland). Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidzorg (Protocol for the execution of economic evaluation in healthcare) [Internet]. 29-02-2016. 2016. p. 120. Available from: https://www.ispor.org/PEguidelines/source/NL-Economic_Evaluation_Guidelines.pdf

16. Lancaster KJ. A new approach to consumer theory. J Polit Econ. 1966;74(2):132–57.

17. Mcfadden D. Conditional logit analysis of qualitative choice behavior. Front Econom. 194AD;105(42).

18. Thurstone LL. A law of comparative judgment. Psychol Rev. 1927 Jul;34(4):273–86.

19. Soekhai V, De Bekker-Grob EW, Ellis AR, Vass CM. Discrete Choice Experiments in Health Economics: Past, Present and Future. Pharmacoeconomics. 123AD;37:201–26.

20. Trapero-Bertran M, Rodríguez-Martín B, López-Bastida J. What attributes should be included in a discrete choice experiment related to health technologies? A systematic literature review. PLoS One. 2019 Jul 1;14(7).

21. Koopmanschap MA, Stolk EA, Koolman X. Dear policy maker: Have you made up your mind? A discrete choice experiment among policy makers and other health professionals. Int J Technol Assess Health Care. 2010;26(2):198–204.

22. Tappenden P, Brazier J, Ratcliffe J, Chilcott J. A stated preference binary choice experiment to explore NICE decision making. Pharmacoeconomics. 2007;25(8):685–93.

23. Wranik WD, Jakubczyk M, Drachal K. Ranking the Criteria Used in the Appraisal of Drugs for Reimbursement: A Stated Preferences Elicitation With Health Technology Assessment Stakeholders Across Jurisdictional Contexts. Value Heal. 2020 Apr 1;23(4):471–80.

24. Green C, Gerard K. Exploring the social value of health-care interventions: A stated preference discrete choice experiment. Vol. 18, Health Economics. 2009. p. 951–76.

25. Pérez-Troncoso D. A step-by-step guide to design, implement, and analyze a discrete choice experiment. 2020;1–10.

26. Jonker MF, Donkers B, de Bekker-Grob E, Stolk EA. Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments. Heal Econ (United Kingdom). 2019;28(3):350–63.

27. P Bridges JF, Brett Hauber A, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint Analysis Applications in Health-a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force Background to the task force report. JVAL. 2011;14:403–13.

28. RIVM. Zorguitgaven blijven tot 2060 stijgen, gemiddeld met 2,8 procent per jaar [Internet]. 2020 [cited 2021 Feb 12]. Available from: https://www.rivm.nl/nieuws/zorguitgaven-blijven-tot-2060-stijgen-gemiddeld-met-28-procent-per-jaar

29. Reckers-Droog V, van Exel N, Brouwer W. Looking back and moving forward: On the application of proportional shortfall in healthcare priority setting in the Netherlands. Health Policy (New York). 2018;122(6):621–9.

30. Johnson FR, Lancsar E, Marshall D, Kilambi V, Mühlbacher 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 Heal. 2013 Jan 1;16(1):3–13.

31. B O. Sample Size Issues for Conjoint Analysis Chapter 7. In: Pu R, editor. Getting started with Conjoint Analysis: Stragtegies for Product Design and Pricing Research. 2010. p. 57–66.

32. Johnson FR, Yang J-C, Reed SD. The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments. Value Heal. 2019;22(2):157–60.

33. Shen J. Latent class model or mixed logit model? A comparison by transport mode choice data. Appl Econ. 2009;41(22):2915–24.

34. Petrova GI, Mariz S, Aballéa S, Aballea@creativ S, Toumi M, Millier A, et al. Social Preferences for Orphan Drugs: A Discrete Choice Experiment Among the French General Population. Front Med. 2020;7:323.

35. Bourke SM, Plumpton CO, Hughes DA. Societal Preferences for Funding Orphan Drugs in the United Kingdom: An Application of Person Trade-Off and Discrete Choice Experiment Methods. Value Heal. 2018 May 1;21(5):538–46.

36. Flynn TN, Bilger M, Malhotra • Chetna, Finkelstein EA. Are Efficient Designs Used in Discrete Choice Experiments Too Difficult for Some Respondents? A Case Study Eliciting Preferences for End-of-Life Care. Pharmacoeconomics. 2016;34(3):273–84.

37. Mitton C, Smith N, Peacock S, Evoy B, Abelson J. Integrating public input into healthcare priority-setting decisions. Evid Policy A J Res Debate Pract. 2011;7(3):327–43.

# Appendices

## Appendix 1 – Pilot questionnaire (Dutch)

Commentaren tijdens het lezen en invullen van de pilot:

Vragen na de pilot:

1. Is de survey begrijpelijk?
2. Kunt u elk attribuut in uw eigen woorden uitleggen?
    a. Ziektelast ja/nee
    b. Samenstelling gezondheidswinst ja/nee
    c. Gezondheidsverbetering ja/nee
    d. Kosteneffectiviteit ja/nee
    e. Grootte patiëntengroep ja/nee
    f. Gemiddelde leeftijd patiëntengroep ja/nee
    g. Vergoeding in vergelijkbare landen ja/nee
3. Wat vindt u van de (lengte van de) uitleg van de attributen
4. Zijn de gekozen levels van de attributen logisch gekozen in uw optiek?
5. Wat vind u van de verschillen tussen de levels? Zijn deze groot genoeg om uw voorkeur aan te geven? (voorbeeld: ziektelast, klein, gemiddeld, groot)
6. Welke manier van kosteneffectiviteit beschrijven had uw voorkeur, en waarom?

Optie A:

- Slechte kosteneffectiviteit, (hoge kosten vergeleken met het verkregen effect)
- Gemiddelde kosteneffectiviteit, (kosten die passen bij het verkregen effect)
- Goede kosteneffectiviteit, (lage kosten vergeleken met het verkregen effect)

Optie B:

- Slechte waarde voor geld
- Gemiddelde waarde voor geld
- Hoge ware voor geld

7. Rangschik de attributen op volgorde van belangrijk naar minst belangrijk

| Attribuut | Ranking |
|---|---|
| Ziektelast | |
| Samenstelling gezondheidswinst | |
| Gezondheidsverbetering | |
| Kosteneffectiviteit | |
| Grootte patiëntengroep | |
| Gemiddelde leeftijd patiëntengroep | |
| Vergoeding in vergelijkbare landen | |

8. Wat vind u van de optie om geen van beide scenario's te kiezen?
9. Wat vond u van de uiteindelijke keuzetaak die gebruikt is in de pilot?
10. Heeft u nog op of aanmerkingen?

## Appendix 2 – Explanation of the attributes (Dutch)

De geneesmiddelscenario's in de keuzetaken bevatten meerdere beoordelingscriteria met verschillende levels. Het is de bedoeling dat u in elke keuzetaak het geneesmiddelscenario kiest dat uw voorkeur heeft om te vergoeden in Nederland. De beoordelingscriteria en levels worden nu uitgelegd.

**Ziektelast**
Dit criteria bekijkt of patiënten ernstig belemmerd worden door hun ziekte. Een manier om de ziektelast uit te drukken is via de kwaliteit van leven op een schaal van 0 – 1 waar 0 dood aangeeft en 1 een volledig goede gezondheid. In deze survey wordt er onderscheid gemaakt tussen drie verschillende ziektelasten:

• Kleine ziektelast, bijvoorbeeld **hypertensie** (score 0.97)
• Gemiddelde ziektelast, bijvoorbeeld **COPD** (score 0.47)
• Grote ziektelast, bijvoorbeeld **een ernstige vorm van de ziekte van alzheimer**, (score - 0.20)

**Gezondheidsverbetering door de behandeling**
Een medicijnbehandeling kan leiden tot een grote gezondheidsverbetering, maar dat hoeft niet. Het effect van een behandeling kan ook erg klein zijn. De grootte van het effect wordt in onderstaande categorieën verdeeld:

• Kleine verbetering
• Gemiddelde verbetering
• Grote verbetering

**Samenstelling van de gezondheidsverbetering**
Naast de grootte van de gezondheidsverbetering is ook de samenstelling van de gezondheidsverbetering van belang bij een behandeling. Een medicijn kan de kwaliteit van leven verbeteren, de duur van het leven verlengen, of beide. In deze survey maken we onderscheid tussen drie soorten gezondheidsverbeteringen:

• 100% Verlenging van leven, 0% kwaliteit van leven
• 100% Verbeterde kwaliteit van leven, 0% verlenging van leven
• 50% Verlenging van leven en 50% verbeterde kwaliteit van leven

Nu volgt de uitleg van de laatste drie beoordelingscriteria die gebruikt worden in deze survey.

**Kosteneffectiviteit**
Dit criterium beschrijft de combinatie van de kosten en effectiviteit van een geneesmiddel. Het beschikbare budget voor geneesmiddelen is eindig en kan maar één keer worden uitgegeven. Hoge kosten nemen een groot deel van het budget in beslag waardoor andere geneesmiddelen mogelijk niet vergoed kunnen worden. We onderscheiden kosteneffectiviteit in drie klasses in deze survey:

• Slechte kosteneffectiviteit, **(hoge kosten vergeleken met het verkregen effect)**
• Gemiddelde kosteneffectiviteit, **(kosten die passen bij het verkregen effect)**
• Goede kosteneffectiviteit, **(lage kosten vergeleken met het verkregen effect)**

**Grootte van de patiëntengroep**
Sommige aandoeningen komen voor bij veel patiënten, andere aandoeningen zijn juist erg zeldzaam. De jaarprevalentie van een aandoening geeft het aantal patiënten weer met de betreffende aandoening binnen één jaar. In deze survey wordt er onderscheid gemaakt tussen drie verschillende groottes:

• Klein, bijvoorbeeld **aids en hiv infecties, jaarprevalentie 26.400**
• Gemiddeld, bijvoorbeeld **ADHD, jaarprevalentie 209.000**
• Groot, bijvoorbeeld **astma, jaarprevalentie 636.200**

**Gemiddelde leeftijd van de patiëntengroep**
Sommige aandoeningen komen vooral voor bij ouderen, anderen juist bij kinderen. In deze survey wordt er onderscheid gemaakt tussen drie leeftijdscategorieën:

• Kinderen **(0 – 18 jaar)**
• Volwassenen **(18 – 67 jaar)**
• Ouderen **( > 67 jaar)**

## Appendix 3 - Syntax and priors of the first and second experimental design

Fist design

```
design
;alts   = altA*, altB*
;eff    = (mnl,d,mean)
;bdraws = halton(300)
;rows   = 24
;block  = 2

;model:
U(altA) = b1.dummy  [(u,0.03,0.10)|(u,0.10,0.17)]        * disease_sev     [2, 3, 1]
        + b2.dummy  [(u,0.07,0.13)|(u,0.13,0.20)]        * health_improv   [2, 3, 1]
        + b3.dummy  [(u,-0.03,0.0)|(u,-0.02,0.02)]       * compos_improv   [2, 3, 1]
        + b4.dummy  [(u,0.07,0.13)|(u,0.13,0.20)]        * cost_effect     [2, 3, 1]
        + b5.dummy  [(u,-0.03,0.03)|(u,-0.03,0.03)]      * patient_pop     [2, 3, 1]
        + b6.dummy  [(u,0.0,0.03)|(u,-0.02,0.02)]        * age             [2, 3, 1]

/
U(altB) = b1.dummy * disease_sev
        + b2.dummy * health_improv
        + b3.dummy * compos_improv
        + b4.dummy * cost_effect
        + b5.dummy * patient_pop
        + b6.dummy * age

$
```

Second design:

```
design
;alts   = altA*, altB*
;eff    = (mnl,d,mean)
;bdraws = halton(300)
;rows   = 24
;block  = 2

;model:
U(altA) = b1.dummy  [(n,0.99,0.14)|(n,1.43,1.04)]        * disease_sev     [2, 3, 1]
        + b2.dummy  [(n,0.83,0.07)|(n,1.72,0.34)]        * health_improv   [2, 3, 1]
        + b3.dummy  [(u,-0.68,0.18)|(u,-0.45,0.25)]      * compos_improv   [2, 3, 1]
        + b4.dummy  [(n,1.45,0.28)|(n,1.94,1.04)]        * cost_effect     [2, 3, 1]
        + b5.dummy  [(n,0.81,0.52)|(n,1.23,0.08)]        * patient_pop     [2, 3, 1]
        + b6.dummy  [(u,-0.52,0.14)|(n,-1.77,0.84)]      * age             [2, 3, 1]

/
U(altB) = b1.dummy * disease_sev
        + b2.dummy * health_improv
        + b3.dummy * compos_improv
        + b4.dummy * cost_effect
        + b5.dummy * patient_pop
        + b6.dummy * age

$
```

## Appendix 4 - Latent class model

This appendix contains the results of the latent class model with 2 classes and a description of these results. The coefficients are listed in table 1. The demographic variables divided over the classes are listed in table 2. Some differences can be seen in the division of demographic variables. The second class contains 16%-points more master students and 13.2%-points more students with a medical background compared with the first class. The differences in age and time till completion were small between the two classes.

Class 1 has a relatively high coefficient for elderly people and has lower coefficients in general compared to class 2. This might be due to factors such as a higher indifference between the attributes or a lower consistency. Population size is ranked as the second most important attribute in class 1. Moderate and high disease severity were not in consequent order in class 1. This might be a sign of the small sample size of the class or that the DCE was not well understood. It is unlikely that people prefer drugs for lower disease severities. In class 2, disease severity is the most important attribute, followed by cost effectiveness and health improvement. In contrast to class 1, class 2 preferers attributes that are in line with the current reimbursement process and places less importance on the attributes that are not explicitly used in the current process. This is in line with the division of demographic variables between the classes where class 2 contains more higher educated students and medical students.

Class 2 shows signs that were in line with the expectations and partly reflect the current HTA process in the Netherlands. Class 1 shows alternative preferences and is less concerned about widely used criteria such as disease severity and cost effectiveness.

*Table 1. Results of the latent class model with two classes.*

| Attribute | Attribute levels | Class 1 | Class 2 |
|---|---|---|---|
| Disease severity | Small disease severity | Reference | |
| | Moderate disease severity | 0.146 | 1.455 |
| | High disease severity | 0.056 | 2.333 |
| Size of the health improvement | Small health improvement | Reference | |
| | Moderate health improvement | 0.502 | 1.142 |
| | Large health improvement | 0.870 | 1.950 |
| Composition of health gain | 50% EoL, 50% QoL | Reference | |
| | 100% EoL | 0.731 | -0.252 |
| | 100% QoL | 0.195 | 0.359 |
| Cost-effectiveness | Bad cost-effectiveness | Reference | |
| | Moderate cost-effectiveness | 0.496 | 1.399 |
| | Good cost-effectiveness | 0.597 | 2.140 |
| Population size | Small population size | Reference | |
| | Moderate population size | 0.090 | 0.633 |
| | Large population size | 0.929 | 1.059 |
| Age of the treatment population | Children | Reference | |
| | Adults | -0.092 | -0.442 |
| | Elderly | -1.897 | -1.202 |
| Class shares | | 0.386 | 0.614 |

*EoL Extension of life, QoL Quality of life.*

*Table 2. Demographic information for the classes of the Latent class model.*

| | | Total | (%) | Class 1 | (%) | Class 2 | (%) |
|---|---|---|---|---|---|---|---|
| Total | | 115 | 100.0% | 45 | 39.1% | 70 | 60.9% |
| Age | | 22.9 | | 22.8 | | 22.9 | |
| Male | | 47 | 40.9% | 17 | 37.8% | 30 | 42.9% |
| Female | | 68 | 59.1% | 28 | 62.2% | 40 | 57.1% |
| Education level | University master | 70 | 60.9% | 23 | 51.1% | 47 | 67.1% |
| | Other | 45 | 39.1% | 22 | 48.9% | 23 | 32.9% |
| Education background | Medical | 68 | 59.1% | 23 | 51.1% | 45 | 64.3% |
| | Other | 47 | 40.9% | 22 | 48.9% | 25 | 35.7% |
| completion time (min) | | 20.8 | | 19.8 | | 21.4 | |

## Appendix 5 – Sensitivity analysis excluding speeders

Table 3 shows the results of the sensitivity analysis that excluded speeders. The coefficients changed only very little and there were no differences in significance. A small change was seen in the heterogeneity after excluding speeders. The p-value of the SD for adults changed from < 0.01 to < 0.1. resulting in a change in significance for this level.

*Table 3. Sensitivity analysis of the main mixed logit model excluding speeders. Standard errors in parentheses.*

| Attribute | Attribute levels | Coefficient (SE) | SD (SE) |
|---|---|---|---|
| Disease severity | Small disease severity | Reference | |
| | Moderate disease severity | 1.007[1] (0.142) | 0.176 (0.286) |
| | High disease severity | 1.564[1] (0.203) | 0.980[1] (0.198) |
| Size of the health improvement | Small health improvement | Reference | |
| | Moderate health improvement | 1.007[1] (0.146) | 0.0709 (0.483) |
| | Large health improvement | 1.685[1] (0.206) | 0.349 (0.293) |
| Composition of health gain | 50% EoL, 50% QoL | Reference | |
| | 100% EoL | -0.0893 (0.165) | 0.986[1] (0.196) |
| | 100% QoL | 0.221 (0.134) | 0.503[2] (0.229) |
| Cost-effectiveness | Bad cost-effectiveness | Reference | |
| | Moderate cost-effectiveness | 1.322[1] (0.160) | 0.0840 (0.232) |
| | Good cost-effectiveness | 1.697[1] (0.212) | 0.894[1] (0.184) |
| Population size | Small population size | Reference | |
| | Moderate population size | 0.633[1] (0.130) | 0.183 (0.433) |
| | Large population size | 1.152[1] (0.148) | 0.0874 (0.436) |
| Age of the treatment population | Children | Reference | |
| | Adults | -0.475[1] (0.131) | 0.409[3] (0.240) |
| | Elderly | -1.820[1] (0.208) | 1.065[1] (0.217) |

*[1] p<0.01, [2] p<0.05, [3] p<0.1, SD standard deviation, SE standard error, QoL quality of life, EoL extension of life.*