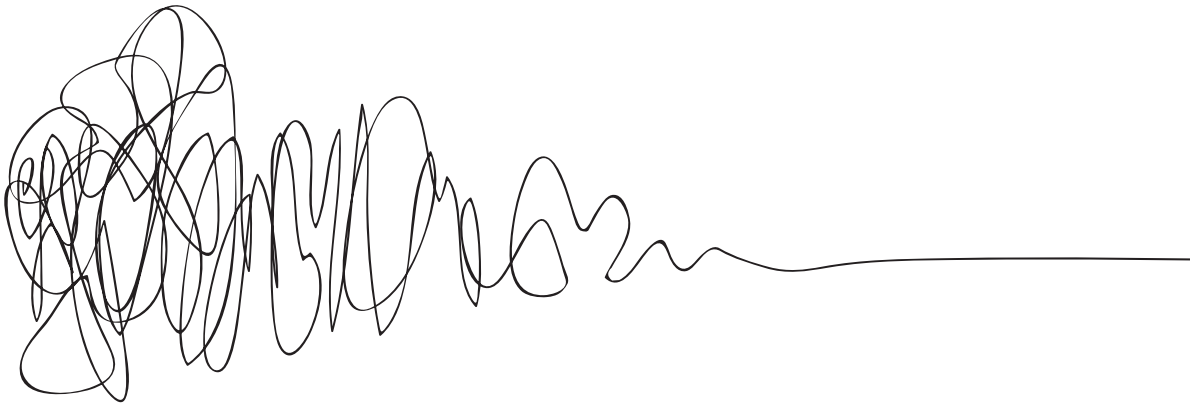
METHODOLOGICAL QUESTIONS REGARDING THE ELICITATION OF PATIENT PREFERENCES AND EMPIRICAL EVIDENCE TO HELP ADDRESS THEM

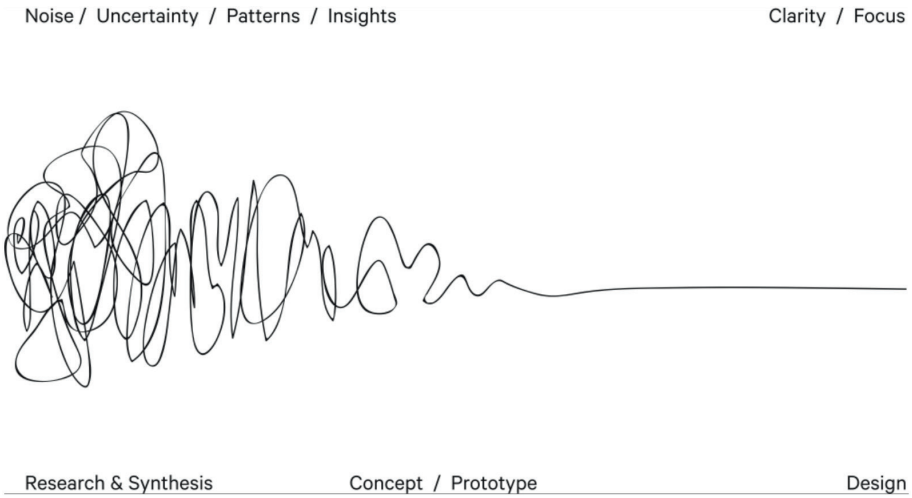


Ian P. Smith

Methodological Questions Regarding the Elicitation of Patient Preferences and Empirical Evidence to Help Address Them

Ian P. Smith

Dit proefschrift werd mogelijk gemaakt met financiële steun van het Patient Preference in Benefit-Risk Assessments during the Drug Life Cycle (PREFER) project en het Innovative Medicines Initiative 2 Joint Undertaking (Grant agreement No. 115966).



"The Squiggle originated as a simple illustration of the design process. The journey of researching, uncovering insights, generating creative concepts, iteration of prototypes and eventually concluding in one single designed solution. It is intended to convey the feeling of the journey. Beginning on the left with mess and uncertainty and ending on the right in a single point of focus: the design." - Damien Newman
"Hey, I recognize this process!" – Ian P. Smith

ISBN: 978-94-6483-844-2
DOI: https://doi.org/10.33540/2060
Cover design: Ridderprint | www.ridderprint.nl
Lay-out: Ridderprint | www.ridderprint.nl
Print: Ridderprint | www.ridderprint.nl

Methodological Questions Regarding the Elicitation of Patient Preferences and Empirical Evidence to Help Address Them

Evidence-based inzichten over methodologische vraagstukken rond het meten van patienten preferenties
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 5 maart 2024 des middags te 2.15 uur

door

Ian Patrick Smith
geboren op 25 juni 1981
te New Haven, CT, Verenigde Staten

Promotor:

Prof. dr. G.A. de Wit

Copromotor:

Dr. J. Veldwijk

Beoordelingscommissie:

Prof. dr. N.J.A. van Exel
Prof. dr. ir. W.G. Goettsch (voorzitter)
Prof. A.V.A. Janssens
Prof. dr. P.G.M. Mol
Prof. dr. D.T.D. de Ridder

TABLE OF CONTENTS

Chapter 1

Introduction

PATIENT INVOLVEMENT IN MODERN MEDICAL PRODUCT DEVELOPMENT AND REGULATION

During much of the 20th century, patients were seen as a passive group who should acquiesce to professional expertise and simply accept medical care as prescribed to them by experts [1, 2]. These viewpoints would hold until societal views changed towards more self-deterministic attitudes of care in the second half of the 20th century [3]. Along with this shift in care came a shift in the role that patients played in developing treatments for their illnesses. Early regulations for drug development had been established to protect patients by ensuring that medical products were safe [4] and effective [5]. However, these regulations had the unintended consequence of increasing costs for drug development which led pharmaceutical companies to focus on medical products with a larger patient base and greater business potential. Patients with less common diseases were "orphaned" as the development of treatments for these diseases were stopped as the costs of development were no longer justified by the limited business potential [6]. In response, patients with these diseases and patient organizations representing rare diseases banded together to lobby regulators for new regulations in the hope that their needs would be addressed [7]. The resulting legislation, the Orphan Drug Act of 1983, provided incentives for pharmaceutical companies to invest in drug development for rare disease by reducing the research costs and burden of evidence needed for approval, speeding up the regulatory review process, and allowing for longer market protection [8]. This regulatory success for patients is largely viewed as a major turning point in the history of patient involvement in drug development and regulation with the patient voice now playing an important role. But it was only the start.

The 1980's also saw large movements of patient advocacy in response to the AIDS epidemic in which activists argued that patients should have a say in what level of risk is acceptable for treatment or participation in clinical trials, not just clinical experts [9]. This advocacy led to the US Food and Drug Administration (FDA) drafting regulations allowing desperately ill patients access to promising new therapies at earlier stages in development [10], and led to internal changes including offices within the FDA who had the specific goal of building relationships with patient communities and initiatives to include patient representatives on advisory committees [11]. Since then, patients and patient advocates have played a larger and larger role in the development and evaluation of medical products with professional stakeholders (i.e. clinicians, pharmaceutical companies, regulators) calling for greater involvement of patients and inclusion of the patient perspective in medical product development [12, 13].

PATIENT PREFERENCES

One specific aspect of the patient perspective that is increasingly being utilized in drug development is the assessment of patient preferences for their care. Patient preferences are "qualitative or quantitative assessments of the relative desirability or acceptability to patients

of specified alternatives or choices among outcomes or other attributes that differ among alternative health interventions" [14]. In simpler terms, the field of patient preference assessment aims to identify what patients want or would accept in their care. The need for this type of research is highlighted in studies showing that the preferences of patients are not always well understood even by those working in the health field [15]. Research of patient preferences can generally be split into two different categories: qualitative and quantitative research [16]. The aim of qualitative or exploratory patient preference research is focused on identifying what is important to patients in their care. This is often done through interviews and focus groups with the aim of compiling aspects of care that are relevant to patients such as which treatment characteristics patients find most relevant [17], exploring unmet medical needs [18], or which factors play a role when deciding whether or not to pursue care [19]. In some instances these exploratory studies are also used to help understand why patients have these preferences such as when looking at why people believe patients with terminal illnesses should or should not be prioritized for care expenses [20] or why video is an important aspect in tele-mental health therapy [21]. While insightful on its own, a primary use of qualitative research in the field of patient preference assessment is to inform the development of tools to quantitatively measure these preferences by identifying what attributes of care patients should be asked about [22, 23].

Where the aim of qualitative preference assessment is to measure what attributes are important to patients, the aim of quantitative patient preference research is to measure how important these attributes are [16]. This information is collected using preference elicitation tasks in which patients need to rank, rate, balance, or choose between different medical products or their characteristics [24]. A common technique to collect this information is a Discrete Choice Experiment [25], but other techniques are also becoming more commonly used such as Best-Worst Scaling (types 1, 2, and 3) [26], Probabilistic Threshold Techniques [27], Time Trade-off [28], Swing Weighting [29], and Analytical Hierarchy Process [30] (to name a few [24]). Examples of their use include identifying priorities for asthma control [31], establishing willingness to pay for improvement in blood glucose control [32], or calculating levels of acceptable risk to reduce symptoms of Parkinson's disease [33].

FOUNDATION AND APPLICATION OF PATIENT PREFERENCES TO THE MEDICAL PRODUCT LIFECYCLE

This thesis will focus on the application of patient preferences in the medical product lifecycle (MPLC). The MPLC is the process of creating a new medical product from initial conception through development, regulation, use in patient care, and post-marketing safety monitoring (see Figure 1)[34]. The foundations for using patient preferences in the MPLC are twofold. First, they are a response to the societal shift towards the rights of the individual patient and the moral belief that every person should be able to choose the path that their life will follow [35]. In the context of medical product development, there is an ethical responsibility to

consider patient values when developing new products and allowing new products onto the market as patients are the ones who will ultimately use these products and bear the burden of the side effects of the treatments. Second, patient preferences can be used to optimize decision making around medical product development to ensure that resources are efficiently allocated, that patient needs are addressed, and that patients are offered healthcare options they see as relevant [34, 36]. Industry stakeholders want to ensure that they are pursuing promising leads and investing funds in products that may someday reach the patient market [37]. Regulatory and reimbursement authorities can apply this information to benefit-risk assessments in order to ground these assessments in patient values [38]. Health technology assessment and reimbursement agencies may use this information to prioritize assets for appraisal and understand the full impact of treatments on the patient when assessing cost-effectiveness [39].

Figure 1. The stages of the medical product lifecycle

Over the previous decades, the field of patient preference research has grown exponentially with hundreds of studies published and the number growing every year [25]. Greater awareness of the use and impact of patient preference studies along with regulatory calls for their inclusion in product assessments has led to greater interest in incorporating of patient preferences at all stages of the MPLC [13, 36]. However, questions remain about how best to incorporate patient preferences into the MPLC and whether the studies being conducted are methodologically sound to ensure that the decisions being informed with this information are well supported [40]. In response to this recognized need, the Innovative Medicine Initiative (IMI) PREFER project was launched in 2017 [40]. The IMI-PREFER project aimed to strengthen patient-centric decision making throughout the MPLC by developing evidence-based recommendations to fill the gaps in knowledge regarding the methodological aspects of patient preference studies. This thesis presents research that was conducted as a part of the IMI-PREFER project.

AIMS AND ORGANIZATION OF THIS THESIS

The aim of this thesis is twofold. First, this thesis aims to identify methodological questions regarding the assessment of patient preferences that are of concern to stakeholders who would potentially use this information in decision-making along the MPLC. Second, this thesis will address several of these questions by presenting research exploring methodological questions and topics identified as being high priority issues.

The first aim is addressed in Chapter 1 which reports on a survey conducted with members of the PREFER consortium. In this survey the stakeholders ranked a list of methodological questions regarding their priority for research to better understand and conduct patient preference studies and support their use in decision making. From this prioritized list, the following research topics will be addressed in this thesis:

1. How generalizable are preferences from one specific population in a disease to different populations in that or related diseases? This research question is addressed in Chapters 2 and 3. **Chapter 2** will compare the preferences of Diabetes patients in the Netherlands and Poland for glucose monitoring technology. **Chapter 3** will present a study identifying common opinions about the most important areas of unmet medical needs in patients with two types of neuro-musculoskeletal diseases, myotonic dystrophy type 1 and mitochondrial disorders and their caregivers.
2. What is the impact of attribute framing on preferences? **Chapter 4** will present the outcomes of a case study in which participants were asked to complete a discrete choice experiment which varied in the way that the attributes were framed (positively, negatively, or both) when presented to participants.
3. How do results differ between simpler/cheaper methods versus more complex/expensive methods? The question of comparability of preference outcomes when measured using different methods (Discrete Choice Experiments and Swing Weighting) will be explored in two case studies. In **Chapter 5** these methods will be compared in a case study assessing patient preferences for glucose monitoring for self-management of diabetes, and in **Chapter 6** they will be compared in a case study measuring treatment preferences of patients with non-small cell lung cancer.
4. How do results differ when participants are presented with information in a video-based educational tool versus traditional text-based education? In **Chapter 7** the results of a randomized study will be presented which directly compares the use of traditional text-based educational material to video-based educational material in a case study in diabetes patients.

Finally, the initial aim of this thesis will be revisited. **Chapter 8** will report on a second survey (conducted 4 years after the study covered in Chapter 1), asking stakeholders in the research-area of patient preference assessment what methodological topics and questions for future studies they think are important to answer, to increase acceptance of preference methods and the use of their results by decision makers. This thesis will then conclude with a summary of the findings, a general discussion of the research conducted, and the implications for future patient preference studies.

REFERENCES

1. European Medicines Agency. *The patient's voice in the evaluation of medicines: how patients can contribute to assessment of benefit and risk.* 2013 18 October 2013; Available from: https://www.ema.europa.eu/en/documents/report/report-workshop-patients-voice-evaluation-medicines_en.pdf.
2. Steele, D.J., et al., *The activated patient: dogma, dream, or desideratum?* Patient education and counseling, 1987. **10**(1): p. 3-23.
3. Roberts, K.J., *Patient empowerment in the United States: a critical commentary.* Health Expectations, 1999. **2**(2): p. 82-92.
4. Fornasier, G., et al., *An historical overview over Pharmacovigilance.* Int J Clin Pharm, 2018. **40**(4): p. 744-747.
5. Greene, J.A. and S.H. Podolsky, *Reform, regulation, and pharmaceuticals--the Kefauver-Harris Amendments at 50.* N Engl J Med, 2012. **367**(16): p. 1481-3.
6. Haffner, M.E., J. Whitley, and M. Moses, *Two decades of orphan product development.* Nat Rev Drug Discov, 2002. **1**(10): p. 821-5.
7. Dunkle, M., *30-year retrospective: National Organization for Rare Disorders, the Orphan Drug Act, and the role of rare disease patient advocacy groups.* Orphan Drugs: Research & Reviews, 2014. **4**: p. 19-27.
8. Herder, M., *What Is the Purpose of the Orphan Drug Act?* PLoS Med, 2017. **14**(1): p. e1002191.
9. Evans, D., *An activist's argument that participant values should guide risk–benefit ratio calculations in HIV cure research.* Journal of Medical Ethics, 2017. **43**(2): p. 100-103.
10. Dunbar, M.M., *Shaking Up the Status Quo: How AIDS Activists Have Challenged Drug Development and Approval Procedures.* Food, Drug, Cosmetic Law Journal, 1991. **46**(5): p. 673-706.
11. Food and Drug Administration. *The History of FDA's Role in Preventing the Spread of HIV/AIDS Washington: Food and Drug Administration*. 2019 14/03/2019 [cited 2023 07/02/2023]; Available from: https://www.fda.gov/about-fda/fda-history-exhibits/history-fdas-role-preventing-spread-hivaids.
12. Brennan, P.F. and I. Strombom, *Improving health care by understanding patient preferences: the role of computer technology.* J Am Med Inform Assoc, 1998. **5**(3): p. 257-62.
13. Jackson, Y., et al., *The evolving role of patient preference studies in health-care decision-making, from clinical drug development to clinical care management.* Expert Review of Pharmacoeconomics & Outcomes Research, 2019. **19**(4): p. 383-396.
14. U.S. Food and Drug Administration, *Patient Preference Information - Voluntary Submission, Review in Premarket Approval Applications, Humatitarian Device Exemption Applications, and De Novo Request, amd Inclusion in Decision Summaries and Device Labeling. Guidance for Industry, Food and Drug Administration Staff and Other Stakeholders*. 2016. p. 29.
15. Rogers, W.A., *Whose autonomy? Which choice? A study of GPs' attitudes towards patient autonomy in the management of low back pain.* Family Practice, 2002. **19**(2): p. 140-145.
16. Soekhai, V., et al., *Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review.* Drug Discovery Today, 2019. **24**(7): p. 1324-1331.
17. Petrocchi, S., et al., *What Matters Most to Lung Cancer Patients? A Qualitative Study in Italy and Belgium to Investigate Patient Preferences.* Frontiers in Pharmacology, 2021. **12**.
18. Jimenez-Moreno, A.C., et al., *Patient Preferences in Rare Diseases: A Qualitative Study in Neuromuscular Disorders to Inform a Quantitative Preference Study.* The Patient - Patient-Centered Outcomes Research, 2021. **14**(5): p. 601-612.

19. Simons, G., et al., *Exploring preferences of at-risk individuals for preventive treatments for rheumatoid arthritis.* Scandinavian Journal of Rheumatology, 2022: p. 1-11.

20. Wouters, S., et al., *Priority to End of Life Treatments? Views of the Public in the Netherlands.* Value in Health, 2017. **20**(1): p. 107-117.

21. Chen, P.V., et al., *Seeing the Value of Video: A Qualitative Study on Patient Preference for Using Video in a Veteran Affairs Telemental Health Program Evaluation.* Telemedicine Reports, 2021. **2**(1): p. 156-162.

22. Coast, J. and S. Horrocks, *Developing attributes and levels for discrete choice experiments using qualitative methods.* Journal of health services research & policy, 2007. **12**(1): p. 25-30.

23. Kløjgaard, M.E., M. Bech, and R. Søgaard, *Designing a stated choice experiment: the value of a qualitative process.* Journal of Choice Modelling, 2012. **5**(2): p. 1-18.

24. Whichello, C., et al., *Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison.* BMC Medical Informatics and Decision Making, 2020. **20**(1): p. 114.

25. Soekhai, V., et al., *Discrete Choice Experiments in Health Economics: Past, Present and Future.* Pharmacoeconomics, 2019. **37**(2): p. 201-226.

26. Flynn, T.N., et al., *Best–worst scaling: What it can do for health care research and how to do it.* Journal of Health Economics, 2007. **26**(1): p. 171-189.

27. Kopec, J.A., et al., *Probabilistic threshold technique showed that patients' preferences for specific trade-offs between pain relief and each side effect of treatment in osteoarthritis varied.* J Clin Epidemiol, 2007. **60**(9): p. 929-38.

28. Lugnér, A.K. and P.F.M. Krabbe, *An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health.* Expert Rev Pharmacoecon Outcomes Res, 2020. **20**(4): p. 331-342.

29. Tervonen, T., et al., *MCDA swing weighting and discrete choice experiments for elicitation of patient benefit-risk preferences: a critical assessment.* Pharmacoepidemiology and Drug Safety, 2017. **26**(12): p. 1483-1491.

30. Danner, M. and A. Gerber-Grote, *Analytic Hierarchy Process*, in *Patient Involvement in Health Technology Assessment*, K.M. Facey, H. Ploug Hansen, and A.N.V. Single, Editors. 2017, Springer Singapore: Singapore. p. 135-147.

31. Ungar, W.J., et al., *Quantifying preferences for asthma control in parents and adolescents using best–worst scaling.* Respiratory Medicine, 2014. **108**(6): p. 842-851.

32. Janssen, E.M., et al., *Education and patient preferences for treating type 2 diabetes: a stratified discrete-choice experiment.* Patient Preference and Adherence, 2017. **11**: p. 1729-1736.

33. Hauber, B., et al., *Parkinson's Patients' Tolerance for Risk and Willingness to Wait for Potential Benefits of Novel Neurostimulation Devices: A Patient-Centered Threshold Technique Study.* MDM Policy Pract, 2021. **6**(1): p. 2381468320978407.

34. van Overbeeke, E., et al., *Factors and situations influencing the value of patient preference studies along the medical product lifecycle: a literature review.* Drug Discov Today, 2019. **24**(1): p. 57-68.

35. Jonsen, A.R., M. Siegler, and W.J. Winslade, *Preferences of Patients*, in *Clinical Ethics: A Practical Approach to Ethical Decisions in Clinical Medicine, 8e*. 2015, McGraw-Hill Education: New York, NY.

36. Whichello, C., et al., *An overview of critical decision-points in the medical product lifecycle: Where to include patient preference information in the decision-making process?* Health Policy, 2020. **124**(12): p. 1325-1332.

37. Paul, S.M., et al., *How to improve R&D productivity: the pharmaceutical industry's grand challenge.* Nature Reviews Drug Discovery, 2010. **9**(3): p. 203-214.

38. van Til, J.A. and M.J. Ijzerman, *Why Should Regulators Consider Using Patient Preferences in Benefit-risk Assessment?* PharmacoEconomics, 2014. **32**(1): p. 1-4.

39. Wale, J.L., et al., *Patients and public are important stakeholders in health technology assessment but the level of involvement is low – a call to action.* Research Involvement and Engagement, 2021. **7**(1): p. 1.

40. de Bekker-Grob, E.W., et al., *Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public-Private Project.* Patient, 2017. **10**(3): p. 263-266.

Chapter 2

Methodological Priorities for Patient Preferences Research: Stakeholder Input to the PREFER Public-Private Project

Authors: Ian P. Smith, Rachael L. DiSantostefano, Esther W. de Bekker-Grob, Bennett Levitan, Conny Berlin, Jorien Veldwijk, G. Ardine de Wit

Status: Published

Smith, I.P., DiSantostefano, R.L., de Bekker-Grob, E.W., Levitan, B., Berlin, C., Veldwijk, J. and de Wit, G.A., 2021. Methodological priorities for patient preferences research: Stakeholder input to the PREFER public–private project. The Patient-Patient-Centered Outcomes Research, 14(5), pp.449-453.

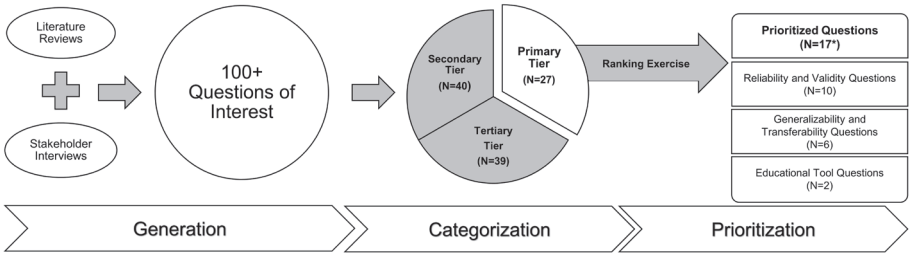
Increasingly, Patient advocacy groups, regulatory agencies, and industry have increasingly advocated for patient engagement in decisions across the medical product lifecycle (MPLC). Amongst the array of approaches to obtain patient input, an increasingly popular and important approach is the use of patient preferences, including an understanding of which endpoints are most important to patients and of the patient perspective on benefit-risk (B-R) tradeoffs when making treatment and reimbursement decisions [1-4]. This push to involve patient preference information throughout the MPLC has resulted in a growing body of knowledge and experience in this field [5-11]. This in turn has stimulated a growing interest in how best to conduct patient preference studies [4, 12-18]. Despite the increasing frequency with which patient preference studies are conducted, there remain many unanswered questions regarding how to incorporate scientifically valid preference measurements into MPLC decision-making regarding medical treatments, including development, regulatory and reimbursement decisions. Previous groups such as the Medical Device Innovation Consortium (MDIC), have worked to address these issues, but many questions remain [15, 19]. To answer some of these questions, the European public-private partnership PREFER ('Patient Preferences in Benefit and Risk Assessments during the Drug Lifecycle') was launched in 2016 [20].

PREFER is a 5-year project, funded jointly by the Innovative Medicines Initiative 2 (EU Horizon 2020) and the European pharmaceutical industry (represented by the European Federation of Pharmaceutical Industries and Associations EFPIA). PREFER aims to strengthen patient-centric decision-making throughout the MPLC by developing evidence-based recommendations to fill the gaps in knowledge regarding methodological aspects of patient preference studies.

The early stages of PREFER focused on assessment of the patient preference landscape to outline research needs. This assessment was based on systematic literature reviews, and both individual interviews and focus groups with key stakeholders (i.e. academic researchers, physicians, regulators, health technology assessment/payer representatives, industry representatives, patients, caregivers and patient representatives). The literature reviews focused on methodological aspects of patient preference assessment methods and previously conducted studies [21-25]. The interviews and focus groups were conducted with more than 140 stakeholders from seven European countries and the United States. These were used to identify the desires, expectations, concerns, and requirements of stakeholders about methodologies for patient-preference elicitation and their use in making well-informed decisions regarding medicinal products [22, 26-29]. Based on the findings of this early work within PREFER, a multi-step approach was used to draft a research agenda for PREFER partners and other parties interested in patient preference information.

The first step in drafting the research agenda involved using the results of the literature reviews and stakeholder interviews to develop over 100 questions on the methodology, design, conduct, and application of preference studies. Irrespective of the point in the medication lifecycle, the

most important research needs identified were related to four high-level concepts: evidentiary standards, assessment of preference heterogeneity, means to minimize patient burden, and means to maximize patient understanding of concepts presented in preference studies. These four high-level concepts were consistent across the literature, methods assessments, interviews and focus groups. The second step involved refining, clustering, and categorizing the questions into three tiers based on their priority and suitability to be examined in prospective case studies within the PREFER project and in subsequent preference studies conducted after the PREFER project has concluded (Figure 1).



*One question was classified under two themes depending on the context of the question

Figure. 1 Process to Determine PREFER Research Agenda and Question Prioritization for Case Studies

The criteria for categorization for each tier were as follows:

- Primary tier: Questions (N=27) relate to the validity and reliability of preference methods, including consistency across preference studies using different methods, adjustments in attributes, and/or different samples. These questions (1) can be examined in a patient preference case study, (2) focus on more promising preference methods [28]; and (3) had not been well studied as of March 2018, according to PREFER partners, stakeholders and external scientific advisors.

- Secondary tier: Questions (N=40) that (1) can be appropriately addressed in a case study, but already have some evidentiary basis, either from previous preference studies or from psychometric research or related disciplines; or (2) relate to topics that are relevant to conducting a preference study such as planning, organization, and set-up.

- Tertiary tier: Questions (N=39) that (1) cannot be appropriately addressed in a case study, or (2) are related to the use and interpretation of patient preference study outcomes.

The classification of research questions into tiers was agreed to by the PREFER consortium and scientific advisory board at the PREFER Annual Meeting in October 2018. After establishing this general research agenda, partners in the PREFER consortium were asked to rank their top

5 questions in the Primary tier according to their priority for being addressed in a prospective case study conducted as a part of PREFER.

In total, 33 members of PREFER partners responded to the survey resulting in the identification of 17 prioritized questions shown in Table 1. The top tier questions relate to three themes: the reliability and validity of preference outcomes, the generalizability and transferability of results, and the impact of educational materials. The reliability and validity questions were generally ranked as the highest priority. These questions involve comparison of different types of preference methods, modulation of specific aspects within a method, or assessment of similar methods across different samples of patients drawn from the same patient population. An example of this type of question is "*How do results differ between simpler/cheaper methods versus more complex/expensive methods?*"

Questions related to generalizability and transferability of patient preference study findings were ranked as a second highest priority. These questions cover topics related to understanding aspects which may explain preference heterogeneity such as differences in recruitment channels, patient characteristics including psychosocial constructs like health literacy and numeracy or locus of control, and variation in preferences across stakeholder groups. This theme was especially highly prioritized by stakeholders who work in clinical settings or directly in patient care. An example of this type of question is "*Can measures of psychosocial constructs serve as covariates that are predictive of preference for particular diseases?*"

Finally, questions related to educational tools used to inform participants in a patient preference studies, and which patient factors to measure to best understand patient preference study outcomes were ranked third most important. An example of these questions is "*How do results differ when participants are presented with information in a scenario-based interactive tool versus traditional text-based education?*"

Based on the final rankings of research questions, each prospective PREFER case study team was encouraged to develop research questions related to these three themes. To this end, the PREFER case studies were asked to include at least one prioritized reliability question, to assess health numeracy and literacy measurements along with other psychosocial constructs, and to address other prioritized questions where possible given the patient population, disease context, and potential design of the case study.

Table 1. Prioritized Questions by Prioritization rank and Question Theme

Prioritization Rank	Question	Question Theme		
		Reliability and Validity	Generalizability and Transferability	Educational Tools
1	How do results differ between simpler/cheaper methods versus more complex/expensive methods?	X		
2	How do changes in the number, type, and definitions of attributes impact results for a given method?	X		
3	How do results differ when different methods with the same set of attributes are applied in the same population?	X		
4	How do results differ when the same method is applied to different samples from the same population?		X	
5	What is the impact of attribute framing on preferences?	X		
6	How generalizable are preferences from one specific population in a disease to different populations in that or related diseases?		X	
7	How to determine which method to use in a given circumstance (and can simulation studies inform this choice)?	X		
8	How do preferences differ when a survey is repeated when an attribute is added or removed?	X		
9	How do results differ when participants are presented with information in a scenario-based interactive tool versus traditional text-based education?			X
10	How to assess whether patients can perform a given set of cognitive tasks?		X	
10	Which attribute presentation formats and combination of formats improves understanding by respondents as shown in increased choice consistency?	X		
10	Which criteria can be used to identify the most suitable preference assessment method to answer a specific preference problem?	X		
13	How can psychosocial constructs be used to explain preference heterogeneity?		X	
14	How does tailoring or personalization of an educational tool based on patient specific characteristics impact a participant's understanding and results?			X
14	To what degree do preferences vary with characteristics of the patients?	X	X	
16	Can measures of psychosocial constructs serve as covariates that are predictive of preference for particular diseases?		X	
16	How do preferences change over time (e.g. as health states and knowledge change)?	X		

The findings of the PREFER project regarding which questions should be prioritized for preference research were developed through expert consensus. These findings corroborated and extended the patient preference research agenda developed by the Medical Device Innovation Consortium (MDIC) whose patient preference report was published the year prior to PREFER's launch [15, 19]. The PREFER project operationalized these agendas by producing a specific set of higher-priority questions to be used to guide the design of prospective case studies. The next steps for PREFER include examining these higher-priority research questions across case studies in three pre-specified patient groups (neuromuscular diseases, lung cancer, and rheumatoid arthritis), nine additional academic and industry led case studies across a wide array of disease and treatment contexts, and simulation studies during the 5-year project.

The questions generated within PREFER cannot all be answered in a fixed number of case studies and the relatively limited timeframe of the PREFER project. Rather, PREFER will advance the field by focusing on topics which require a large public-private collaboration to initiate, and lay the groundwork for future researchers to add to and improve our understanding of patient preferences. To this end, PREFER researchers encourage others to utilize the specific research questions presented in this article to conduct future studies, build off the methodological insights from the PREFER case studies and contribute toward answering the high-prioritized questions. The collective knowledge created by the collaboration will result in a strong body of evidence to help increase the understanding and utilization of patient preference studies across the MPLC.

REFERENCES

1. Soekhai V, De Bekker-Grob EW, Ellis AR, Vass CM. Discrete Choice Experiments in Health Economics: Past, Present and Future. PharmacoEconomics. 2019;37(2):201-26. doi: 10.1007/s40273-018-0734-2.
2. Bridges JFP, Kinter ET, Kidane L, Heinzen RR, McCormick C. Things are Looking up Since We Started Listening to Patients. The Patient: Patient-Centered Outcomes Research. 2008;1(4):273-82. doi: 10.2165/1312067-200801040-00009.
3. Brett Hauber A, Fairchild AO, Reed Johnson F. Quantifying Benefit–Risk Preferences for Medical Interventions: An Overview of a Growing Empirical Literature. Applied Health Economics and Health Policy. 2013;11(4):319-29. doi: 10.1007/s40258-013-0028-y.
4. Marsh K, Van Til JA, Molsen-David E, Juhnke C, Hawken N, Oehrlein EM, et al. Health Preference Research in Europe: A Review of Its Use in Marketing Authorization, Reimbursement, and Pricing Decisions—Report of the ISPOR Stated Preference Research Special Interest Group. Value in Health. 2020;23(7):831-41. doi: 10.1016/j.jval.2019.11.009.
5. Hoos A, Anderson J, Boutin M, Dewulf L, Geissler J, Johnston G, et al. Partnering With Patients in the Development and Lifecycle of Medicines. Therapeutic Innovation & Regulatory Science. 2015;49(6):929-39. doi: 10.1177/2168479015580384.
6. Mavris M, Furia Helms A, Bere N. Engaging patients in medicines regulation: a tale of two agencies. Nature Reviews Drug Discovery. 2019. doi: 10.1038/d41573-019-00164-y.
7. Haerry D, Landgraf C, Warner K, Hunter A, Klingmann I, May M, et al. EUPATI and Patients in Medicines Research and Development: Guidance for Patient Involvement in Regulatory Processes. Frontiers in Medicine. 2018;5. doi: 10.3389/fmed.2018.00230.
8. Anderson M, Kimberly Mccleary K. On the path to a science of patient input. Science Translational Medicine. 2016;8(336):336ps11-ps11. doi: 10.1126/scitranslmed.aaf6730.
9. European Medicines Agency. ICH reflection paper on proposed ICH guideline work to advance patient focused drug Development. In: Committee for Medicinal Products for Human Use, editor.2020.
10. European Medicines Agency. The patient's voice in the evaluation of medicines: How patients can contribute to assessment of benefit and risk. In: Stakeholders and Communication Division, editor.2013.
11. U. S. Food and Drug Administration. Patient-Focused Drug Development: Collecting Comprehensive and Representative Input; Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. In: Center for Drug Evaluation and Research, editor.2020.
12. U. S. Food and Drug Administration. Patient Preference Information – Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling: Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. Silver Spring, MD: U. S. Food and Drug Administration; 2016.
13. Ho M, Saha A, McCleary KK, Levitan B, Christopher S, Zandlo K, et al. A Framework for Incorporating Patient Preferences Regarding Benefits and Risks into Regulatory Assessment of Medical Technologies. Value in Health. 2016;19(6):746-50. doi: 10.1016/j.jval.2016.02.019.
14. Reynolds RF, Kurz X, De Groot MCH, Schlienger RG, Grimaldi-Bensouda L, Tcherny-Lessenot S, et al. The IMI PROTECT project: purpose, organizational structure, and procedures. 2016;25:5-10. doi: 10.1002/pds.3933.

15. Medical Device Innovation Consortium. Medical Device Innovation Consortium (MDIC) Patient Centered Benefit-Risk Project Report: A Framework for Incorporating Information on Patient Preferences Regarding Benefit and Risk into Regulatory Assessments of New Medical Technology. 2015.

16. Postmus D, Richard S, Bere N, Van Valkenhoef G, Galinsky J, Low E, et al. Individual Trade-Offs Between Possible Benefits and Risks of Cancer Treatments: Results from a Stated Preference Study with Patients with Multiple Myeloma. The oncologist. 2018;23(1):44-51. doi: 10.1634/theoncologist.2017-0257.

17. Mühlbacher AC, Juhnke C, Beyer AR, Garner S. Patient-Focused Benefit-Risk Analysis to Inform Regulatory Decisions: The European Union Perspective. Value in Health. 2016;19(6):734-40. doi: 10.1016/j.jval.2016.04.006.

18. Johnson FR, Zhou M. Patient Preferences in Regulatory Benefit-Risk Assessments: A US Perspective. Value in Health. 2016;19(6):741-5. doi: 10.1016/j.jval.2016.04.008.

19. Levitan B, Hauber AB, Damiano MG, Jaffe R, Christopher S. The Ball is in Your Court: Agenda for Research to Advance the Science of Patient Preferences in the Regulatory Review of Medical Devices in the United States. The Patient - Patient-Centered Outcomes Research. 2017;10(5):531-6. doi: 10.1007/s40271-017-0272-6.

20. De Bekker-Grob EW, Berlin C, Levitan B, Raza K, Christoforidi K, Cleemput I, et al. Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public–Private Project. The Patient - Patient-Centered Outcomes Research. 2017;10(3):263-6. doi: 10.1007/s40271-017-0222-3.

21. Russo S, Jongerius C, Faccio F, Pizzoli SFM, Pinto CA, Veldwijk J, et al. Understanding Patients' Preferences: A Systematic Review of Psychological Instruments Used in Patients' Preference and Decision Studies. Value in Health. 2019;22(4):491-501. doi: 10.1016/j.jval.2018.12.007.

22. Van Overbeeke E, Whichello C, Janssens R, Veldwijk J, Cleemput I, Simoens S, et al. Factors and situations influencing the value of patient preference studies along the medical product lifecycle: a literature review. Drug Discovery Today. 2019;24(1):57-68. doi: 10.1016/j.drudis.2018.09.015.

23. Janssens R, Huys I, Van Overbeeke E, Whichello C, Harding S, Kübler J, et al. Opportunities and challenges for the inclusion of patient preferences in the medical product life cycle: a systematic review. BMC Medical Informatics and Decision Making. 2019;19(1). doi: 10.1186/s12911-019-0875-z.

24. Soekhai V, Whichello C, Levitan B, Veldwijk J, Pinto CA, Donkers B, et al. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discovery Today. 2019;24(7):1324-31. doi: 10.1016/j.drudis.2019.05.001.

25. Whichello C, Bywall KS, Mauer J, Stephen W, Cleemput I, Pinto CA, et al. An overview of critical decision-points in the medical product lifecycle: Where to include patient preference information in the decision-making process? Health Policy. 2020. doi: 10.1016/j.healthpol.2020.07.007.

26. Van Overbeeke E, Janssens R, Whichello C, Schölin Bywall K, Sharpe J, Nikolenko N, et al. Design, Conduct, and Use of Patient Preference Studies in the Medical Product Life Cycle: A Multi-Method Study. Frontiers in Pharmacology. 2019;10. doi: 10.3389/fphar.2019.01395.

27. Janssens R, Russo S, Van Overbeeke E, Whichello C, Harding S, Kübler J, et al. Patient Preferences in the Medical Product Life Cycle: What do Stakeholders Think? Semi-Structured Qualitative Interviews in Europe and the USA. The Patient - Patient-Centered Outcomes Research. 2019;12(5):513-26. doi: 10.1007/s40271-019-00367-w.

28. Whichello C, Levitan B, Juhaeri J, Patadia V, Disantostefano R, Pinto CA, et al. Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison. BMC Medical Informatics and Decision Making. 2020;20(1). doi: 10.1186/s12911-020-01142-w.

29. Van Overbeeke E, Forrester V, Simoens S, Huys I. Use of Patient Preferences in Health Technology Assessment: Perspectives of Canadian, Belgian and German HTA Representatives. The Patient - Patient-Centered Outcomes Research. 2020. doi: 10.1007/s40271-020-00449-0.

Chapter 3

Diabetes Patient Preferences for Glucose Monitoring Technologies: Results from a Discrete Choice Experiment in Poland and the Netherlands

Authors: Ian P Smith, Chiara L Whichello, Jorien Veldwijk, Maureen P.M.H. Rutten-van Mölken, C.G.M Groothuis-Oudshoorn, Rimke C. Vos, Esther W de Bekker-Grob, G. Ardine de Wit

Status: Published

Smith, I.P., Whichello, C.L., Veldwijk, J., Rutten-van Mölken, M.P., Groothuis-Oudshoorn, C.G., Vos, R.C., de Bekker-Grob, E.W. and De Wit, G.A., 2023. Diabetes patient preferences for glucose-monitoring technologies: results from a discrete choice experiment in Poland and the Netherlands. BMJ Open Diabetes Research and Care, 11(1), p.e003025.

ABSTRACT

Introduction: New glucose-monitoring technologies have different cost-benefit profiles compared to traditional finger-prick tests resulting in a preference-sensitive situations for patients. This study aimed to assess the relative value adults with diabetes assign to device attributes in two countries.

Research Design and Methods: Adults with type 1 or 2 diabetes from the Netherlands (n=226) and Poland (n=261) completed an online discrete-choice experiment (DCE). Respondents choose between hypothetical glucose monitors described using seven attributes: precision, effort to check, number of finger-pricks required, risk of skin irritation, information provided, alarm function and out-of-pocket costs. Panel mixed-logit models were used to determine attribute relative importance and calculate expected uptake rates and willingness-to-pay.

Results: The most important attribute for both countries was monthly out-of-pocket costs. Polish respondents were more likely than Dutch respondents to choose a glucose monitoring device over a standard finger-prick and had higher willingness-to-pay for a device. Dutch respondents had higher willingness-to-pay for device improvements in effort to check and reducing the number of finger-pricks a device requires.

Conclusions: Costs are the primary concern of patients in both countries when choosing a glucose monitor and would likely hamper real-world uptake. The costs-benefit profiles of such devices should be critically reviewed.

1. INTRODUCTION

Diabetes is a chronic disease characterized by the body's inability to maintain healthy levels of blood glucose which is associated with long-term health problems including an increased risk of mortality with an estimated global prevalence of 10.5% in 2021 [1-3]. Diabetes care is centered around the cornerstone of metabolic control; specifically keeping glucose levels as close to normal as possible through medication, a careful diet, physical activity, and self-monitoring of blood glucose (SMBG) [4, 5]. SMBG has traditionally been done using a finger-prick test and is associated with improvements in glycemic control [6]. While highly accurate [7], this technique represents a large burden to patients which can result in non-compliance to medical treatment advice [8-10]. Studies examining the adherence of patients to SMBG regimens report adherence rates ranging from 88% in Australia [11] to as low as 44% in Sweden [12], 26% in the USA [13], and 20% in Hungary [14]. These low adherence rates are related to barriers to the practice of SMBG including low socio-economic status (SES), fear of testing and fingertip pain, distressing emotions and thoughts, frustration about "poor" blood glucose reading, lack of awareness of hypoglycemia and hyperglycemia, lack of social support, and difficulty in interpreting SMBG results [15].

Recent technological developments have resulted in commercially available medical devices which can (semi-)continuously monitor blood glucose levels (or proxies thereof) [16, 17]. These devices are often less invasive, quicker, and easier to use, and can give more detailed daily blood glucose level information by showing trends over time compared to SMBG with finger-pricking [18-20]. However, these devices vary in regard to functionality and features including (but not limited to) differences in accuracy, size, battery requirements, range of transmitter, calibration requirements, scanning procedures, and longevity (replacement time). Further, these devices are often not reimbursed through insurance plans and can have high out-of-pocket costs for the patient [21]. The differences in function, features, and costs have resulted in a situation where personal preferences may the choice of device for SMBG.

Despite growing interest in patient preference assessment, limited research has been done quantifying patient preferences for glucose monitors. Hannah et al. found that for type 1 diabetes patients (T1DM) the most important factors for choosing a continuous glucose monitor (CGM) were method of data retrieval, longer sensor wear time with more adhesive durability, and personalized alerts and alarms [22]. Engler et al. found that the reasons related to stopping CGM usage for T1DM patients were poor accuracy due to lag times, insurance reimbursement or cost, comfort, and false alarms [18]. They also found that for T1DM patients without CGM experience cost, having a device attached to the body, and expectations of discomfort in wearing were primary reasons for not using a CGM for SMBG [18]. Both studies highlight the preference sensitive nature of these devices, however neither included type 2 diabetes patients (T2DM), a growing population of patients who may need to monitor their blood glucose [19, 20, 23]. Further, only Hannah et al. [22] used a method of relative

valuation to show how important these attributes were in regards to each other but did not include a cost attribute which is a major concern for many patients. There is thus a gap in knowledge regarding the relative valuation information that regulatory authorities and decision-makers use to guide policies for medical treatments [24]. This study aimed to fill that gap by quantitatively assessing the factors that T1DM and T2DM patients consider important when choosing a glucose monitoring device for SMBG and identify willingness-to-pay and expected device uptake rates.

2. MATERIALS AND METHODS

2.1 Subjects

Participants were recruited from the Netherlands and Poland through a professional panel provider (SurveyEngine). These countries were chosen as costs were expected to play an important role in deciding between devices and these two countries had partial and no reimbursement of glucose monitors for SMBG at the time of data collection (respectively). To be eligible to complete the survey patients had to have a self-reported diagnosis of T1DM or T2DM, reside in the Netherlands or Poland, be over 18 years of age, be able to read and understand Dutch or Polish, and have access to a computer.

2.2 Discrete Choice Experiment

A Discrete Choice Experiment (DCE) was used to quantify patient preferences [25, 26]. DCEs are based on Random Utility Theory (RUT) which assumes that the utility or value of a healthcare alternative can be derived through the compound valuation of the different attributes and attribute-levels used to describe the treatment alternative [27-29]. In a DCE, respondents are presented with choice tasks in which they chose their preferred option from two or more alternative treatment profiles. These alternative profiles describe treatments using a set of characteristics (called attributes) with varying levels, representing realistic values of these attributes [30, 31]. Patients choose the alternative which represents the highest personal utility based on the personal value they attach to the different levels of attributes used to describe the alternative. After a patient completes the DCE, attribute estimates can be generated using econometric models and the relative importance of the included attributes can be inferred from these estimates [32-34].

2.2.1 Attributes and Levels

The attributes and levels used in this DCE were developed according to best practices using a stepwise, qualitative approach from April to October 2019 [35, 36]. This approach started with a scoping literature review of articles describing aspects relevant to patients in using glucose-monitoring devices. The results of this review were used to create an interview guide

(see supplementary material Figure 1) which was used in semi-structured interviews with T1DM and T2DM patients from the Netherlands (n=9), clinical diabetes experts (n=5), patient organization representatives (n=2), and pharmaceutical industry representatives involved in glucose monitoring device development (n=4), as well as a focus group with T1DM and T2DM patients in Poland (n=10). This process generated a list of 12 potentially relevant attributes which was reviewed and reduced by the research team to ensure relevance according to the interviewees, non-redundancy, and operationality to a final list of 7 attributes for use in the DCE. The levels used to describe the attributes were developed based on the literature review and interviews and were chosen to be realistic and reflect the most common types of commercially available glucose monitors, including CGMs and flash glucose monitors (FGM) [37-40]. One attribute ('out-of-pocket costs') was standardized between the two countries using purchasing power parity weights to assure that the relative value of the levels was similar given the differences in wealth between the two countries [41]. The final list of attributes and levels used in the DCE can be found in Table 1.

2.2.2 Experimental Design

The DCE was developed using an efficient design (Bayesian D-efficient design [42, 43]) generated in NGene 1.0 software. This allows for participants to complete a minimal amount of choice tasks (three blocks of N=12 choice tasks each) while maximizing the amount of information each task generates. Available literature, interviews, and researcher knowledge were used to generate the initial design. The design was updated after a pilot of N=99 Dutch participants. In each choice task patients were instructed to imagine that their doctor told them to check their blood glucose levels at least four times per day and gave them options of devices to choose from to do this. The choice tasks were presented using a dual

2.2.2 Experimental Design

The DCE was developed using an efficient design (Bayesian D-efficient design [42, 43]) generated in NGene 1.0 software. This allows for participants to complete a minimal amount of choice tasks (three blocks of N=12 choice tasks each) while maximizing the amount of information each task generates. Available literature, interviews, and researcher knowledge were used to generate the initial design.

The design was updated after a pilot of N=99 Dutch participants. In each choice task patients were instructed to imagine that their doctor told them to check their blood glucose levels at least four times per day and gave them options of devices to choose from to do this. The choice tasks were presented using a dual response "best-best" set-up where participants first chose between two hypothetical glucose monitors (Device A or Device B) and then they between that choice and a standard finger-prick test for their care (see Figure 1) [44, 45]. This method mimics realistic choice scenarios while also ensuring data quality. The finger-prick test was

always described as requiring four finger-pricks per day, a high amount of effort to check blood glucose, no skin irritation or redness associated with a device on skin, showing glucose level only at time of measurement, no alarm, and with out-of-pocket costs of €25 (or 55zl) per month. Participants were given two 'warm-up' DCE choice-tasks before the main exercise started to ensure comprehension.

2.3 Questionnaire

Prior to completing the DCE, participants were given information describing glucose monitoring as a part of diabetes self-management including the impact of uncontrolled blood glucose on health outcomes, and a description of the attributes used in the DCE. Participants were asked to answer sociodemographic questions and disease related questions including diabetes type, years since diagnosis, use of medication, and questions related to their current diabetes self-care regimen. Two brief measures assessing subjective numeracy (the Shortened Subjective Numeracy Scale (SNS-3)) [46] and health literacy (Brief Health Literacy Screener (Chew Items)) [47] were included in the survey. The final survey was pre-tested in think-aloud interviews with N=6 diabetes patients from the Netherlands. The outcomes of this pre-test were used to reword the survey for understandability.

Figure 1. Example DCE Choice task

Imagine that your doctor told you to check your blood glucose levels at least four times per day. To do this, the doctor offers you different hypothetical devices to choose from.

	Device A		Device B
Precision compared to fingerpricking	Less accurate than fingerpricking (higher or lower by 0.3)		Less accurate than fingerpricking (higher or lower by 0.6)
Average number of fingerpricks per day	0		0
Effort to check	Low effort		Moderate effort
Probability of getting skin irritation or redness	5% chance of skin irritation or redness (5 out of 100)		35% chance of skin irritation or redness (35 out of 100)
Glucose information	Current Glucose level		Current Glucose level and arrow
Alarms	Yes		No
Monthly costs	€25		€175
I prefer:	❍		❍

If you have to choose between the device you have chosen above and the traditional fingerprick-test to check your glucose levels, which one would you prefer? (Please note that a fingerprick-test should be done four times a day, requires high effort to check, does not result in skin irritation or redness, will show your glucose levels, doesn't have an alarm and costs €25 per month).

Select only one answer

❍ **I prefer the device I have selected above**	❍ **I prefer the fingerprick-test**

Table 1: Attributes and levels for the discrete choice experiment

Attributes	Level 1	Level 2	Level 3	Level 4
Precision compared to finger-pricking [a]	Less accurate than finger-pricking (higher or lower by 0.6 mmol/L (*10.8 mg/dL)	Less accurate than finger-pricking (higher or lower by 0.3 (*5.4 mg/dL)	Accurate as finger-pricking §	---
Average number of finger-pricks per day [b]	4 §	2	0	---
Effort to check [c]	High effort: you need to measure your glucose levels yourself	Moderate effort: you scan a sensor to check glucose levels	Low effort: glucose levels automatically sent to you §	---
Probability of getting skin irritation or redness [d]	35% chance of skin irritation or redness	20% chance of skin irritation or redness	5% chance of skin irritation or redness	No chance of skin irritation or redness §
Monthly costs [e]	€250 (*550zl)	€175 (*390zl)	€100 (*220zl)	€25 (*55zl) §
Glucose information [f]	Current Glucose level §	Current Glucose level and arrow	Current Glucose level and a graphic of your level trends over the day	---
Alarms [g]	No §	Yes	---	---

* Unit equivalents shown for Polish survey

§ Reference level

(a-g) Attribute explanations as presented to patients:

A. Some glucose monitors are more precise than others. Finger-pricking is generally regarded as the most accurate way to measure glucose levels. Measurements from devices that use sensors can be just as accurate, but can also be less accurate than finger-pricking, especially if your glucose levels are very high or very low. For example, if your glucose level is 6 mmol/L and you measure it with a device that is off by 0.6 mmol/L, then this device can say your glucose is anywhere from 5.4 to 6.6 mmol/L

B. This is how many times you would need to do a finger-prick-test each day on an average day. This number could be higher on days when you feel the need to test more often like when you're sick, but we want you to picture an average day. Sometimes, this is your only method of measuring your glucose levels. Or, you might need to do finger-prick-tests to confirm the levels from another device

C. This means how much effort you need to give to check your blood glucose levels. High effort checking means you need to stop what you're doing and concentrate on measuring your levels. You need to wash your hands, get out your device equipment, prick your finger, put blood on a strip, check the results, and then clean everything up. Moderate effort checking means you need to get out a small device and use it to scan the sensor on your body to obtain your glucose levels. Low effort checking means your glucose levels are automatically sent to a device which you can view at any time. This could be a dedicated glucose device, your phone, or a smartwatch. You don't need to do anything to have your blood glucose levels sent through, just look at the device to check.

D. A chance of skin irritation or redness around a sensor means a redness or itchy rash on the skin around or under the sensor. This is similar to having an itchy allergic reaction and can be rather uncomfortable or irritating. The sensor will need to be removed and replaced in a different spot. This skin irritation and redness usually lasts until after the sensor is replaced. Not all sensor have this side effect so chances of getting the side effect can differ per device. If a device gives you a 15% chance, this means that 15 out of a 100 people who get this device experience skin irritation and redness while 85 out of a 100 people do not experience this.

E. This means how much money you need to pay out-of-pocket per month in order to check your blood glucose. Please note that this is money that is not reimbursed by your insurance. This could be money needed to pay for devices, sensors, or strips used.

F. This means how your glucose levels are presented to you. This information could be only your current glucose level (you only see a digital number like 8.3 mmol/L). This could be your current glucose level with an arrow showing how your blood glucose is changing as compared to your previous measurement (increasing, decreasing, stable). Or, it could show your current glucose level with a graphic of your blood glucose levels over the day.

G. Your device will give you a beeping alarm (like a phone notification) any time your blood glucose levels are (getting) too high or too low.

2.4 Analysis

2.4.1 Data Quality

Respondents were required to answer all questions and only surveys that included all necessary questions for the final analysis were included. Completed responses were checked for flat-lining (only choosing Device A or Device B) and speeding (respondents completing the survey faster than 70% of the mean response time based on log data) as data quality checks. Differences in sample demographics were assessed using chi-square tests or t-tests where applicable. A significance of $p<0.05$ was used for all analyses.

2.4.2 Preferences

Data from the DCE was analyzed by combining the two questions from each choice task as one single observation (Device A versus Device B versus the finger-prick-test). Preference estimates in each country were assessed independently using a panel mixed-effects logit regression to account for heterogeneity of preferences within patient populations [32]. Effects-coding was used for all variables except for cost which was assumed to be linear [48]. Effects-coding allows for a calculation of the reference category coefficient which can be used for comparison to other attributes and a clear interpretation of a constant term (reflecting the utility of a status-quo finger-prick test) [48]. Robust outcomes were generated by applying 1,000 Halton draws [49]. The analysis was conducted in STATA version 14 [34]. The optimal model was identified based on log likelihood. Attributes with significant standard deviations for at least one level were included as random effects in the final model. The following value functions were used for the final analyses:

Equations 1-3:

$$V_{Device\ A\ i} = \beta_0 + \beta_{1i} * precision_{0.3} + \beta_{2i} * precision_{0.6} + \beta_{3i} * pricks\ per\ day_{2x} + \beta_{4i} * effort_{moderate} + \beta_{5i} * skin\ irritation_{20\%} + \beta_{6i} * skin\ irritation_{35\%} + \beta_{7i} * monthly\ costs + \beta_{8i} * information_{arrow} + \beta_{9i} * information_{trendline} + \beta_{10i} * alarms_{none}$$

$$V_{Device\ B\ i} = \beta_{1i} * precision_{0.3} + \beta_{2i} * precision_{0.6} + \beta_{3i} * pricks\ per\ day_{2x} + \beta_{4i} * effort_{moderate} + \beta_{5i} * skin\ irritation_{20\%} + \beta_{6i} * skin\ irritation_{35\%} + \beta_{7i} * monthly\ costs + \beta_{8i} * information_{arrow} + \beta_{9i} * information_{trendline} + \beta_{10i} * alarms_{none}$$

$$V_{Fingerprick\ i} = \beta_{11}$$

In these equations, the value of an alternative for individual i is calculated based on the coefficients reflecting the relative importance of each attribute or attribute-level ($\beta 1$ to $\beta 10$). $\beta 11$ is an alternative specific constant reflecting the individual's preference for the fixed alternative of the finger-prick-test over Device B. $\beta 0$ is a constant term which identifies the respondent's preferences for Device A over Device B, reflecting a left-right bias in case participants had a tendency to favor the left option. All attributes and attribute-levels were included as random parameters, with a normal distribution to identify heterogeneity in the preferences for those attributes.

The mixed-logit model preference estimates were used to calculate attribute relative importance scores (RIS) [50]. The RIS reflects how important one attribute is compared to another. These were calculated by identifying the attribute with the greatest absolute difference between highest and lowest valued level and using this as a reference (RIS = 1). The RIS for each attribute were then calculated as the quotient of the absolute difference of the most and least valued level of that attribute and the reference value. This results in a normalized scale for comparison.

2.4.3 Willingness-to-pay estimates and uptake rates

Individual attribute coefficient estimates were extracted from the mixed-effects models to calculate individual willingness-to-pay (WTP) estimates and expected uptake rates. WTP estimates were generated by calculating the utility difference between attribute levels and dividing this by the negative linear cost coefficient resulting in the estimated amount that each participant would be willing to pay for the change in attribute level. Very small cost coefficients for some participants led to extreme WTP outliers so the median and inter-quartile range are reported rather than the mean. Differences in median WTP estimates were assessed using a Mood's Test for equality of medians [51].

Expected uptake rate estimates were calculated using the individual attribute coefficient estimates. Three device profiles represent potential glucose monitoring devices were used to calculate uptake rates compared to a standard finger-prick test. The first profile represented the most desired device according to the outcomes of the mixed-logit model: high precision, zero finger-pricks per day, low effort to check, low chance of skin irritation, €25 per month out-of-pocket costs, glucose information with a daily trendline, and an alarm. The second profile was similar to a generic FGM: moderate precision, zero finger-pricks, moderate effort, moderate chance of skin irritation, €100 per month out-of-pocket costs, glucose information with an arrow indicating glucose direction, and no alarm. The last profile used the generic FGM profile but changed the monthly out-of-pocket costs to €25. The uptake estimate was calculated at the individual level by taking the proportion of the individuals'(i) total utility which was accountable to a device (V) in a scenario containing both this device and a finger-prick alternative (W) using the following equation:

Equation 4: $\sum_{i=1}^{n} \frac{e^{V_i}}{e^{V_i} + e^{Fingerprick_i}}$

The mean of these expected uptake rate estimates was interpreted as the expected population uptake rate.

2.5 Ethics

This study was approved by the Medical Research Ethics Committee of the UMC Utrecht (WAG/mb/19/045208). The study was conducted according to the principles of the Declaration

of Helsinki. All participants were informed about the study through written materials and provided written informed consent prior to participating in the study.

3. RESULTS

In total, N=521 respondents completed the surveys. Of those, N=487 responses were included in the final analysis after N=34 (6.5%) respondents were excluded following a check of data quality. Participant demographic information can be found in Table 2. Compared to the Polish sample, the Dutch sample was significantly older (51.6 years vs 39.4 years), had lived with diabetes for more years, were less educated, had lower levels of health numeracy, and were less likely to monitor their blood glucose than. No other significant differences were found between the samples.

3.1 Preferences for Glucose Monitors

All attributes were found to be significant for patients in at least one of the countries. Significant heterogeneity of preferences was found for all attributes except for type of glucose information. High costs were associated with a lower likelihood of choosing a device. Increased precision was preferred over lower precision, and decreased number of finger-pricks and chance of skin irritation were consistently favored over increases in these attributes. Samples from both countries favored a device with an alarm over one without an alarm. Improving a device's effort to check from moderate to low and improving glucose information to show more than only current levels were only important for the Dutch respondents. Both samples preferred glucose monitoring devices over a finger-prick test. The complete results of the mixed-logit model can be found in Table 3.

Regarding the RIS of the attributes, costs were found to be the most important factor when choosing a device by a factor of five compared to the next most important attribute and a factor of approximately 50 compared to the least important attribute (Supplementary material Figure 2). For the Dutch sample, after costs the most important attributes were number of finger-pricks, followed by precision and chance of skin irritation all of which were comparably valued. For the Polish population, after costs precision of device was the second most important attribute followed by chance of skin irritation. These were also comparably valued. Polish patients were not as averse to additional finger-pricks as Dutch respondents and found this approximately half as important as Dutch respondents. However, Polish respondents valued switching to a device from a finger-prick test more than Dutch respondents. Having an alarm and improving glucose information were both relatively unimportant in a device. Only the Dutch sample viewed improved effort to check and the type of glucose information as important when deciding on a device.

Table 2: Respondent characteristics (n=487)

Characteristics	Dutch respondents N= 226	Polish respondents N=261
Age in years ** (mean ± sd)	51.6 ± 17.2	39.4 ± 13.4
Sex (n, %)		
Females	116 (51.3)	125 (47.9)
Males	110 (48.7)	134 (51.3)
Type of diabetes (n, %)		
Type 1	65 (28.8)	83 (31.8)
Type 2	158 (69.9)	167 (64.0)
Other	3 (1.3)	11 (4.2)
Number of years having diabetes ** (mean ± sd, (median, range))	9.5 ± 9.1 (6.5, 0-60)	6.1 ± 7.1 (3, 0-53)
Current Glucose monitor used as part of diabetes care **		
CGM or FGM	38 (16.8)	39 (14.9)
Finger-prick testing only	128 (56.6)	211 (80.8)
None	60 (26.5)	11 (4.2)
Checks glucose more than 2x per day*	83 (31.8)	161 (71.2)
Uses insulin (n, %)	120 (53.1)	140 (53.6)
Health literacy		
High	102 (45.1)	113 (43.3)
Low	124 (54.9)	148 (56.7)
Numeracy*		
High	195 (86.3)	243 (93.1)
Low	31 (13.7)	18 (6.9)
OECD Educational level ** (n, %)		
Tertiary	100 (44.2)	134 (51.3)
Upper-Secondary/Vocational	114 (50.4)	127 (48.7)
Secondary or Lower	12 (20.8)	0 (0.0)

* Significant differences between countries at p<0.05; ** Significant differences between countries at p<0.001; CGM: continuous glucose monitor; FGM: flash glucose monitor; OECD: Organisation for Economic Co-operation and Development

Table 3: Attribute-level estimates for the panel mixed-logit model

			Netherlands			Poland		
Attribute	**Levels**		**Estimate**	**S.E.**	**p sig.**	**Estimate**	**S.E.**	**p sig.**
Precision compared to Finger-pricking	**Accurate as finger-pricking (ref)**	Mean	0.343	0.075	***	0.457	0.071	***
		S.D.						
	±0.3 mmol/L	Mean	0.000	0.061		-0.081	0.051	
		S.D.	0.036	0.101		0.047	0.112	
	±0.6 mmol/L	Mean	-0.343	0.079	***	-0.376	0.073	***
		S.D.	0.536	0.093	***	0.655	0.082	***
Average number of finger-pricks per day	**0 times per day (ref)**	Mean	0.352	0.059	***	0.172	0.044	***
		S.D.						
	2 times per day	Mean	-0.352	0.059	***	-0.172	0.044	***
		S.D.	0.479	0.070	***	0.349	0.059	***
Effort to check	**Low (ref)**	Mean	0.120	0.039	**	0.042	0.033	
		S.D.						
	Moderate	Mean	-0.120	0.039	**	-0.042	0.033	
		S.D.						
Probability of getting skin irritation or redness	**5% (ref)**	Mean	0.336	0.076	***	0.377	0.064	***
		S.D.						
	20%	Mean	-0.059	0.066		-0.018	0.059	
		S.D.	0.061	0.127		0.015	0.166	
	35%	Mean	-0.277	0.076	***	-0.359	0.066	
		S.D.	0.450	0.097	***	0.402	0.084	***
Monthly costs	**per €1 increase**	Mean	-0.017	0.002	***	-0.016	0.001	***
		S.D.	0.015	0.001	***	0.019	0.001	***
Glucose information	**Current glucose level only (ref)**	Mean	-0.142	0.063	*	-0.056	0.054	
		S.D.						
	Current glucose level with Arrow	Mean	0.068	0.063		0.004	0.055	
		S.D.						
	Current glucose level with Daily Trendline	Mean	0.074	0.063		0.052	0.053	
		S.D.						
Alarms	**Yes (ref)**	Mean	0.152	0.044	***	0.148	0.035	***
		S.D.						
	No	Mean	-0.152	0.044	***	-0.148	0.035	***
		S.D.	0.252	0.076	***	0.151	0.063	*
Alternative specific constant for device instead of finger-prick-test		Mean	-0.982	0.502		-2.770	0.336	***
		S.D.	4.527	0.386	***	4.767	0.371	***
Alternative specific constant indicating left-right bias		Mean	0.376	0.085	***	0.346	0.074	***
		S.D.	0.446	0.140	***	0.540	0.098	***

Higher estimates represent increasing levels of importance for the patient in choosing a device;
* indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$;
S.D. = standard deviation; ref = reference level;
Note: All attributes were effects-coded, enabling the direct comparison of the estimates. The sum of the effect coded attributes is zero, and therefore the coefficient of the reference category can be easily calculated and the relative importance of the reference categories of the attributes can be compared with one another, and so that the alternative specific constants have independent interpretation signifying the average utility for that alternative. S.D.'s are given where parameters were found to have a significant random parameter estimate. The significant alternative specific constant indicates that patients tended to choose the alternative on the left side. A normal distribution using 1000 Halton draws was used in model development.

3.2 Willingness-to-pay for a Glucose Monitor and Expected Uptake Rates

WTP results can be found in Table 4. It was estimated that Polish patients would pay significantly more to switch from standard finger-prick to a device than Dutch patients (€65.01 vs €27.74 per month). The median WTP for improvements in glucose monitors ranged from €2.58 (for the Dutch respondents to improve glucose information) to €33.64 (for the Polish respondents to increase precision from low to high). Significant differences were found between the two countries with Dutch respondents having higher WTP for device improvements in precision from low to medium, improving effort to check, and improving glucose information. Dutch patients were also willing to pay significantly more for a reduction in number of finger-pricks per day in conjunction with a device compared to Poland (€32.71 vs €13.35).

These differences were also reflected in the expected uptake rates for devices. Polish patients were significantly more likely to choose a device over finger-prick (Table 4) compared to Dutch patients. These differences were most pronounced in patients aged 18-50, patients with T2DM, and current finger-prick only users.

4. DISCUSSION

To the best of our knowledge, this was the first study to investigate the relative importance of different attributes describing glucose monitoring technologies which involved cost as an attribute. As expected, cost was found to be the most important factor for patients when deciding on glucose monitors in both the Netherlands and Poland. Increased device precision, reduction in skin irritation, and required number of finger-pricks per day were the next most important attributes when choosing between glucose monitors.

The findings from this study replicate some of the findings of earlier studies [18, 22], but the current study enables us to show that costs were at least more than five times more important for patients when choosing a glucose monitor than any other attribute. As costs are the primary consideration for patients when deciding to use a glucose monitoring device or a standard finger-prick, it may not be a question of willingness to pay, but ability to pay that is determining glucose monitor choice [18, 52-54]. This is unfortunate as the improvements in diabetes outcomes, patient quality of life, and healthcare expenditures in connection with using these devices are increasingly documented [17, 54-60].

Beyond costs, the relative importance of the other attributes differed to some degree between the two countries. Specifically, Dutch respondents valued reducing the number of daily finger-pricks to zero more than twice as much as Polish respondents. The acceptance of additional finger pricks to verify blood glucose levels may reflect the greater importance that Polish respondents assigned to precision as these finger-pricks are the most accurate reading and can be used for calibration of devices or verification of device glucose information. For both populations, precision was mainly significant when the device was described as having higher

Table 4. Median WTP estimates and Average Uptake rates compared to traditional finger-prick

	Netherlands WTP		Poland WTP		
	Median	**IQR**	**Median**	**IQR**	
Increase precision from:					
Low to medium	15.94	(27.73, 3.93)	10.18	(-3.19, 21.23)	***
Medium to high	15.87	(3.72, 26.43)	22.17	(4.54, 41.84)	
Low to high	31.82	(7.60, 54.73)	33.64	(1.02, 59.45)	
Reduce daily finger-pricks: 2 to 0	32.71	(14.34, 63.41)	13.35	(4.28, 30.34)	***
Improve chance of skin irritation:					
High to medium	10.39	(18.86, 5.56)	13.80	(6.38, 27.87)	**
Medium to low	18.70	(12.52, 32.79)	16.06	(8.20, 34.52)	*
High to low	28.97	(16.55, 52.5)	29.83	(14.47, 61.74)	
Improve effort from medium to low	11.32	(8.78, 22.24)	3.55	(2.31, 7.4)	***
Improve glucose information with:					
An arrow showing blood glucose is changing	10.22	(7.70, 19.49)	2.58	(1.68, 5.38)	***
Daily trend information	14.19	(7.92, 20.07)	4.60	(2.99, 9.59)	***
Get a glucose alarm	14.19	(7.15, 25.74)	12.67	(7.44, 24.02)	
Willingness to pay to not use finger-prick test	27.74	(-231.85, 278.23)	65.01	(-183.76, 295.5)	**

	Most preferred device (%) [a]		**FGM proxy device (%) [b]**		**FGM proxy with reduced cost (%) [c]**	
Total samples						***
Netherlands (n=226)	63.6	***	44.4	**	54.8	
Poland (n=261)	77.1		56.1		67.6	
Age 18-50					59.9	*
Netherlands (n=88)	69.1	*	51.7	***	69.5	
Poland (n=202)	78.9		59.5			
Age 50 and over						
Netherlands (n=137)	60.4		39.9		51.9	
Poland (n=59)	70.9		44.3		61.1	
FP only users						***
Netherlands (n=128)	57.5	***	37.8	***	48.0	
Poland (n=211)	75.0		53.6		65.8	
CGM/FGM users					73.0	
Netherlands (n=38)	78.2		66.7		76.7	
Poland (n=39)	85.3		70.2			
Type 1						
Netherlands (n=65)	70.8		53.1		62.0	
Poland (n=83)	81.7		63.2		71.7	
Type 2						
Netherlands (n=157)	61.0	**	41.0	**	52.3	**
Poland (n=167)	74.8		52.8		65.6	

WTP: Willingness-to-pay

Significant differences between countries: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Note for Willingness-to-pay estimates: estimates are only presented for attribute improvement where level increases were found to be significant in the mixed logit model; Costs presented in euros; Mood's test of equality of medians was used to assess difference between countries.

a High precision, 0 finger-pricks, low effort, low chance of skin irritation, €25/month, glucose information with trendline, alarm

b Moderate precision, 0 finger-pricks, moderate effort, moderate chance of skin irritation, €100/month, arrow information, no alarm

c Moderate precision, 0 finger-pricks, moderate effort, moderate chance of skin irritation, €25/month, arrow information, no alarm

levels of imprecision. Lower levels of imprecision were not important for choosing a device indicating that there is an acceptable amount of device imprecision. This was also reported by patients during the qualitative phase.

These preference differences resulted in different in willingness-to-pay for glucose devices and expected uptake rates for the two countries. Both samples reflected an overall desire to move away from finger-prick tests for SMBG although this was more pronounced in the Polish population compared to the Dutch population. Patients were consistently willing to pay for device improvements that resulted in devices that more closely represented FGMs or CGMs regarding functionality.

While we found type of information to be relatively less important based on the model outcomes, this conflicts with the findings from the qualitative phase of this study. During the interviews, stakeholders from every area including the patients indicated that only having the current glucose level was insufficient for proper glucose management. In the preference study outcomes, improvements in this area were not nearly as important for choosing a device as the interviews would have led us to believe. In addition to this, industry interviewees and patients reported that connectivity to devices which the patients normally carried around (e.g., smartphone, smartwatch) was a desirable feature as it reduces effort to check and the stigma of checking blood glucose levels. The preference outcomes indicated that while the Dutch patients significantly preferred a device with low burden, the added benefit of accessing this information on a smartphone or watch instead of a dedicated device was relatively limited compared to other features or costs. This indicates that connectivity is something that is a want but not a must in a device. Exceptions to this may be in specific instances, such as parents who want to be able to monitor a child's glucose level at a distance [61, 62].

Our case study focused on two countries, the Netherlands and Poland, which are examples of 'Western' and 'Eastern' European countries with partial and no reimbursement for glucose monitoring devices supporting the transferability of these findings to other countries with out-of-pocket costs for SMBG. At the time of designing the study, the reimbursement for CGMs was limited in the Netherlands with FGMs not fully reimbursed [21]. The reimbursement policy of Dutch insurance companies changed while the study was being conducted to allow T1DM patients, T2DM patients with intense insulin regimens, and T2DM patients who are pregnant or trying to become pregnant to be eligible for FGMs through their health insurance. CGMs and FGMs were not reimbursed in Poland at the time that the study was conducted and to the best of our knowledge are still not reimbursed [63, 64]. Respondent awareness of the change in reimbursement in the Netherlands may have resulted in lower WTP estimates. It would be interesting to study how improved access to these devices for some patients has changed preferences in Dutch patient populations and if the removal of cost as an attribute impacts their preferences compared to Poland without a change in reimbursement. The removal of cost as a barrier would likely have a large impact on patient preferences and expected uptake

rates of these devices with a greater focus on how the device fits into the patient's lifestyle as reflected in the study by Hannah et al [22].

The strengths of this study include the extensive qualitative phase used to identify the relevant attributes for use in the DCE. This process was more extensive than what is commonly done to generate attributes in preference studies. Interviewees were internationally diverse with a broad range of backgrounds and contributed to the identification of a set of attributes relevant to a broad sample of patients. Another strength of this study is the multi-country sampling which allows for a better understanding of the transferability of these findings to diabetes patient populations in other countries. This study did have some limitations. First, the study the collected data relied on self-reports of diabetes diagnosis and no quotas based on SES were imposed. This limited exploring sub-group analyses of SES group preferences which may be relevant as SES has previously been associated with adherence to SMBG. Second, patients were recruited through an online panel only and not through clinical partners or patient organizations due to COVID-19 related restrictions on all non-vital research. This resulted in a sample of respondents that had generally higher levels of education and were younger than we would expect from the general diabetes population [65-68]. The results of a more representative sample may produce different relative preference outcomes as we found differences in expected uptake rates based on age stratifications.

5. CONCLUSION

While patients value many aspects of glucose monitors, out-of-pocket costs are the primary concern of patients when deciding on devices to self-monitor blood glucose. Even when different welfare levels between the two countries were accounted for, differences in estimated willingness-to-pay were found between the countries. This study shows that uptake of modern glucose monitoring devices is dependent on out-of-pocket costs. In light of these clear preferences to switch from glucose measurement by finger-pricks to more modern equipment, a critical review of the costs and benefits of such devices is needed to see if removing the cost barrier is justified by the potential improvements in blood glucose monitoring.

REFERENCES

1. American Diabetes Association, Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, 2012. 35(Supplement 1): p. S64-S71.
2. Sun, H., et al., IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice, 2022. 183: p. 109119.
3. Saeedi, P., et al., Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice, 2020. 162: p. 108086.
4. Evert, A.B., et al., Nutrition Therapy Recommendations for the Management of Adults With Diabetes. Diabetes Care, 2014. 37(Supplement_1): p. S120-S143.
5. Rubin, R.R. and M. Peyrot, Psychological issues and treatments for people with diabetes. Journal of Clinical Psychology, 2001. 57(4): p. 457-478.
6. Canadian Agency for Drugs Technologies in Health, Systematic review of use of blood glucose test strips for the management of diabetes mellitus. CADTH technology overviews, 2010. 1(2).
7. Benjamin, E.M., Self-Monitoring of Blood Glucose: The Basics. Clinical Diabetes, 2002. 20(1): p. 45-47.
8. Ahola, A.J. and P.H. Groop, Barriers to self-management of diabetes. Diabetic Medicine, 2013. 30(4): p. 413-420.
9. Dalewitz, J., N. Khan, and C.O. Hershey, Barriers to Control of Blood Glucose in Diabetes Mellitus. American Journal of Medical Quality, 2000. 15(1): p. 16-25.
10. Peyrot, M., et al., Psychosocial problems and barriers to improved diabetes management: results of the Cross-National Diabetes Attitudes, Wishes and Needs (DAWN) Study. Diabetic Medicine, 2005. 22(10): p. 1379-1385.
11. Bruce, D.G., et al., Diabetes education and knowledge in patients with type 2 diabetes from the community. Journal of Diabetes and its Complications, 2003. 17(2): p. 82-89.
12. Moström, P., et al., Adherence of self-monitoring of blood glucose in persons with type 1 diabetes in Sweden. BMJ Open Diabetes Research and Care, 2017. 5(1): p. e000342.
13. Harris, M.I., C.C. Cowie, and L.J. Howie, Self-monitoring of blood glucose by adults with diabetes in the United States population. Diabetes care, 1993. 16(8): p. 1116-1123.
14. Hankó, B., et al., Self-reported medication and lifestyle adherence in Hungarian patients with type 2 diabetes. Pharmacy World & Science, 2007. 29(2): p. 58-66.
15. Ong, W.M., S.S. Chua, and C.J. Ng, Barriers and facilitators to self-monitoring of blood glucose in people with type 2 diabetes using insulin: a qualitative study. Patient preference and adherence, 2014. 8: p. 237.
16. Garg, S.K. and H.K. Akturk, The future of continuous glucose monitoring. 2017, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.
17. Ajjan, R.A., How Can We Realize the Clinical Benefits of Continuous Glucose Monitoring? Diabetes Technology & Therapeutics, 2017. 19(S2): p. S-27-S-36.
18. Engler, R., T.L. Routh, and J.Y. Lucisano, Adoption Barriers for Continuous Glucose Monitoring and Their Potential Reduction With a Fully Implanted System: Results From Patient Preference Surveys. Clinical Diabetes, 2018. 36(1): p. 50-58.

19. Carlson, A.L., D.M. Mullen, and R.M. Bergenstal, Clinical Use of Continuous Glucose Monitoring in Adults with Type 2 Diabetes. Diabetes Technology & Therapeutics, 2017. 19(S2): p. S-4-S-11.

20. Taylor, P.J., C.H. Thompson, and G.D. Brinkworth, Effectiveness and acceptability of continuous glucose monitoring for type 2 diabetes management: A narrative review. Journal of Diabetes Investigation, 2018. 9(4): p. 713-725.

21. Heinemann, L., et al., Reimbursement for Continuous Glucose Monitoring: A European View. Journal of Diabetes Science and Technology, 2012. 6(6): p. 1498-1502.

22. Hannah, K., et al., PDB40 Eliciting Patient Preferences for Continuous Glucose Monitoring Devices in Type 1 Diabetes. Value in Health, 2021. 24: p. S85.

23. Harris, M.I., Frequency of Blood Glucose Monitoring in Relation to Glycemic Control in Patients With Type 2 Diabetes. Diabetes Care, 2001. 24(6): p. 979-982.

24. Zhang, Y., et al., Using patient values and preferences to inform the importance of health outcomes in practice guideline development following the GRADE approach. Health and Quality of Life Outcomes, 2017. 15(1).

25. De Bekker-Grob, E.W., M. Ryan, and K. Gerard, Discrete choice experiments in health economics: a review of the literature. Health Economics, 2012. 21(2): p. 145-172.

26. Soekhai, V., et al., Discrete Choice Experiments in Health Economics: Past, Present and Future. PharmacoEconomics, 2019. 37(2): p. 201-226.

27. Lancaster, K.J., A New Approach to Consumer Theory. Journal of Political Economy, 1966. 74(2): p. 132-157.

28. Manski, C.F., The structure of random utility models. Theory and Decision, 1977. 8(3): p. 229-254.

29. McFadden, D., Econometric models of probabilistic choice. Structural analysis of discrete data with econometric applications, 1981. 198272.

30. Gerard, K., M. Shanahan, and J. Louviere, Using Discrete Choice Modelling to Investigate Breast Screening Participation, in Using Discrete Choice Experiments to Value Health and Health Care, M. Ryan, K. Gerard, and M. Amaya-Amaya, Editors. 2008, Springer Netherlands: Dordrecht. p. 117-137.

31. Ryan, M., K. Gerard, and M. Amaya-Amaya, Using discrete choice experiments to value health and health care. Vol. 11. 2007: Springer Science & Business Media.

32. Hensher, D.A. and W.H. Greene, The Mixed Logit model: The state of practice. Transportation, 2003. 30(2): p. 133-176.

33. Hensher, D.A., et al., Applied choice analysis: a primer. 2005: Cambridge university press.

34. Louviere, J.J., D.A. Hensher, and J.D. Swait, Stated choice methods: analysis and applications. 2000: Cambridge university press.

35. Reed Johnson, F., et al., Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value in Health, 2013. 16(1): p. 3-13.

36. Bridges, J.F.P., et al., Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health, 2011. 14(4): p. 403-413.

37. (FDA), F.a.D.A., Blood Glucose Monitoring Test Systems for Prescription Point-of-Care Use. 2018, Food and Drug Administration (FDA).

38. Carlson, A.L., D.M. Mullen, and R.M. Bergenstal, Clinical use of continuous glucose monitoring in adults with type 2 diabetes. Diabetes Technology and Therapeutics, 2017.

39. Heinemann, L. and J.H. DeVries, Reimbursement for Continuous Glucose Monitoring. Diabetes Technology and Therapeutics, 2016.

40. Allen, N.A., et al., Continuous glucose monitoring in non-insulin-using individuals with type 2 diabetes: Acceptability, feasibility, and teaching opportunities. Diabetes Technology and Therapeutics, 2009.

41. OECD. OECD (2022), Purchasing power parities (PPP) (indicator). doi: 10.1787/1290ee5a-en (Accessed on 10 February 2022). 2022 [cited 2019 September].

42. Kanninen, B.J., Optimal Design for Multinomial Choice Experiments. Journal of Marketing Research, 2002. 39(2): p. 214-227.

43. Rose, J.M. and M.C.J. Bliemer, Constructing Efficient Stated Choice Experimental Designs. Transport Reviews, 2009. 29(5): p. 587-617.

44. Ghijben, P., E. Lancsar, and S. Zavarsek, Preferences for Oral Anticoagulants in Atrial Fibrillation: a Best–Best Discrete Choice Experiment. PharmacoEconomics, 2014. 32(11): p. 1115-1127.

45. Veldwijk, J., et al., The Effect of Including an Opt-Out Option in Discrete Choice Experiments. PLoS ONE, 2014. 9(11): p. e111805.

46. Fagerlin, A., et al., Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. Medical Decision Making, 2007. 27(5): p. 672-680.

47. Chew, L.D., K.A. Bradley, and E.J. Boyko, Brief questions to identify patients with inadequate health literacy. health, 2004. 11: p. 12.

48. Bech, M. and D. Gyrd-Hansen, Effects coding in discrete choice experiments. Health Economics, 2005. 14(10): p. 1079-1083.

49. Train, K.E., Drawing from Densities, in Discrete Choice Methods with Simulation, K.E. Train, Editor. 2009, Cambridge University Press: Cambridge. p. 205-236.

50. Gonzalez, J.M., A Guide to Measuring and Interpreting Attribute Importance. The Patient - Patient-Centered Outcomes Research, 2019. 12(3): p. 287-295.

51. Mood, A.M., Introduction to the Theory of Statistics. 1950.

52. Zgibor, J.C. and D. Simmons, Barriers to Blood Glucose Monitoring in a Multiethnic Community. Diabetes Care, 2002. 25(10): p. 1772-1777.

53. Divan, V., et al., Perceived Burdens and Benefits Associated With Continuous Glucose Monitor Use in Type 1 Diabetes Across the Lifespan. Journal of Diabetes Science and Technology, 2022. 16(1): p. 88-96.

54. Halford, J. and C. Harris, Determining Clinical and Psychological Benefits and Barriers with Continuous Glucose Monitoring Therapy. Diabetes Technology & Therapeutics, 2010. 12(3): p. 201-205.

55. Bonora, B., et al., Head-to-head comparison between flash and continuous glucose monitoring systems in outpatients with type 1 diabetes. Journal of Endocrinological Investigation, 2016. 39(12): p. 1391-1399.

56. Dover, A.R., et al., Flash Glucose Monitoring Improves Outcomes in a Type 1 Diabetes Clinic. Journal of Diabetes Science and Technology, 2016. 11(2): p. 442-443.

57. Oyagüez, I., et al., Cost analysis of the flash monitoring system (FreeStyle Libre 2) in adults with type 1 diabetes mellitus. BMJ Open Diabetes Research & Care, 2020. 8(1): p. e001330.

58. Leelarathna, L. and E.G. Wilmot, Flash forward: a review of flash glucose monitoring. Diabetic Medicine, 2018. 35(4): p. 472-482.

59. Kelly, B., FreeStyle Libre use in a real-world population: The Southampton City experience. Journal of Diabetes Nursing, 2019. 23(3).

60. Leinung, M., S. Thompson, and E. Nardacci, Benefits of Continuous Glucose Monitor Use in Clinical Practice. Endocrine Practice, 2010. 16(3): p. 371-375.

61. Elbalshy, M., et al., Exploring Parental Experiences of Using a Do-It-Yourself Solution for Continuous Glucose Monitoring Among Children and Adolescents With Type 1 Diabetes: A Qualitative Study. Journal of Diabetes Science and Technology, 2020. 14(5): p. 844-853.

62. Gammon, D., et al., Parent-Child Interaction Using a Mobile and Wireless System for Blood Glucose Monitoring. Journal of Medical Internet Research, 2005. 7(5): p. e57.

63. Hohendorff, J., et al., Intermittently Scanned Continuous Glucose Monitoring Data of Polish Patients from Real-Life Conditions: More Scanning and Better Glycemic Control Compared to Worldwide Data. Diabetes technology & therapeutics, 2021. 23(8): p. 577-585.

64. Głowińska-Olszewska, B., et al., Monthly use of a real-time continuous glucose monitoring system as an educational and motivational tool for poorly controlled type 1 diabetes adolescents. Advances in medical sciences, 2013. 58(2): p. 344-352.

65. OECD, Education at a Glance: OECD Indicators Poland. 2019.

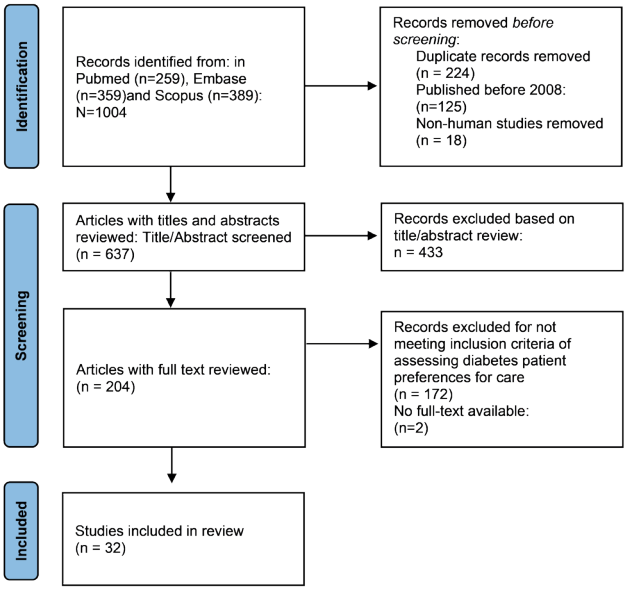
66. Rijksinstituut voor Volksgezondheid en Milieu. Diabetes mellitus: Opleiding. 2021 [cited 2021 01/19/2021]; Available from: https://www.vzinfo.nl/diabetes-mellitus#node-prevalentie-diabetes-huisartsenpraktijk-naar-opleiding.

67. Rijksinstituut voor Volksgezondheid en Milieu. Diabetes mellitus: Leeftijd en geslacht. 2021 [cited 2021 01/19]; Available from: https://www.vzinfo.nl/diabetes-mellitus/leeftijd-en-geslacht.

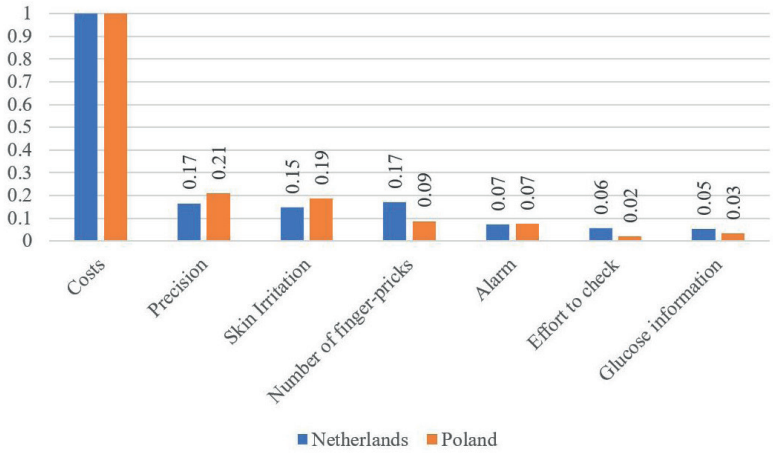
68. Polakowska, M. and W. Piotrowski, Incidence of diabetes in the Polish population. Results of the Multicenter Polish Population Health Status Study – WOBASZ. Polish Archives of Internal Medicine, 2011. 121(5): p. 156-163.

SUPPLEMENTARY MATERIALS

Appendix Figure A1. PRISMA 2020 flow diagram for scoping review to identify relevant attributes related to choice of glucose measurement equipment



Appendix Figure A2. Comparison of Relative Importance Scores of attributes, Netherlands versus Poland



Note: Costs had the highest absolute difference between highest and lowest level coefficients (NL=4.15, PO=3.95) and was used to standardize the other attributes

Focus group guidelines - FOR CLINICIANS, PATIENT ORGANISATION REPS, INDUSTRY REPS

Black = Speech to be directed to participants
[Italics] = Guidance for moderators or instructions
Numbers = Timing indications, how long each section should take

Instructions to moderators:

- *Create warm and friendly environment*
- *Interact with participants, and stimulate interaction between participants*
- *Make seating arrangements for participants according to their needs*
- *Exercise mild unobtrusive control (moderate the discussion but do not interrupt too often)*
- *Adequate knowledge of topic*
- *Have the discipline of listening and apply active listening*
- *Take into account the different types of participants and try to balance the conversation while addressing the obligatory topics: dominant talkers, shy participants, etc*

00.00 Welcome

Welcome, my name is…… and I will be your moderator guiding the discussion. Today we would like to discuss your thoughts about glucose monitoring devices, and any opinions, preferences, or concerns you might have. This discussion is part of a large European project called PREFER, which aims to make patients more involved in the development of their drugs or medical devices.The opinions collected today will be used to write reports and articles to inform companies, health authorities, and other researchers about what matters most to patients when it comes to their choices about glucose monitoring.

There are no right or wrong answers, we are looking for your personal opinions. It is possible that you might not agree with each other, but it would be nice if you could listen respectfully to each other. We're audio recording, so we'd like to ask that only one person is speaking at a time. Also please put your mobile phones on silent. We will use first names today, but your names will not be used in any reports. If there are any questions or terms that are used during the focus groups that are not clear to you, please let us know.

00.05

So, we will now start the focus group. Is that OK for everybody?
Let's go around the circle and say your first name and where you work

00.12

Today we'll be talking about glucose monitoring devices. There are lots of different kinds of devices that can help you monitor your glucose.

- Firstly, how do you think patients feel about using a device to monitor their glucose?

00.15

There are several different kinds of glucose measuring devices, as you can see in your handout, including finger-prick devices and new devices becoming more available called continuous glucose monitors. These are small devices that have a tiny sensor that's inserted under the skin. This is attached to a transmitter that sends your blood sugar levels to a hand-held display device for you to look at, or even to your insulin pump if you have one. The device measures your glucose levels throughout the day and night, and lets you see trends over time, and can give you alerts if you are having high or low blood sugar. There is information and a picture in your participant information sheet

- What are your opinions about continuous glucose monitors?
- How do continuous glucose monitoring devices compare to devices that use a drop of blood from your finger?
- What do you think about when making a comparison?

00.25

Continuous glucose monitors would let you see your glucose levels over a long period of time. They can also help you tell you if your glucose is currently rising or dropping (and at what speed) helping you with timing your insulin, diet, and exercise

- How important is this to patients?

00.30

These devices have a sensor that needs to go under the skin.

- How do you think patients feel about that?
- How do you think patients feel about how it looks?
- How do you think patients feel about its size?

00.35

The sensor needs to be replaced every so often, depending on what kind of model and brand it is. Some sensors need to be replaced every two weeks, other every six months. With some

models, patients can insert the sensor into their skin, by themselves, at home. With other models, they need a doctor.

- How difficult to patients find it to replace the sensor themselves?
- Are there any challenges you can foresee with replacing sensors?
- How important do you think it is for patients that the sensor can have a long life?

00.55

Continuous glucose monitors can give you an alert or alarm to warn you when your glucose is getting too high or too low

- How important do you think this might be to patients?
- When might this be convenient?
- When might this be inconvenient?

01.05

Some continuous glucose monitors can be linked to an insulin pump, and automatically control your insulin depending on how high or low your blood glucose is

- How important do you think this might be to patients?
- When might this be convenient?
- When might this be inconvenient?

01.10

Some continuous glucose monitors still need you to check your blood glucose twice a day using a finger-prick test, to make sure it's measuring your glucose accurately. This is called calibration. There are other devices that don't need you to do this finger-prick check at all. But they are sometimes less accurate when your glucose is very low.

- Which do you think would be more preferable to patients?
- Which would be in their best interest?

01.20

Final questions now

- What do you think is most important to patients when choosing a new glucose monitoring device?
- What do you think is least important to patients when choosing a new glucose monitoring device?

01.30

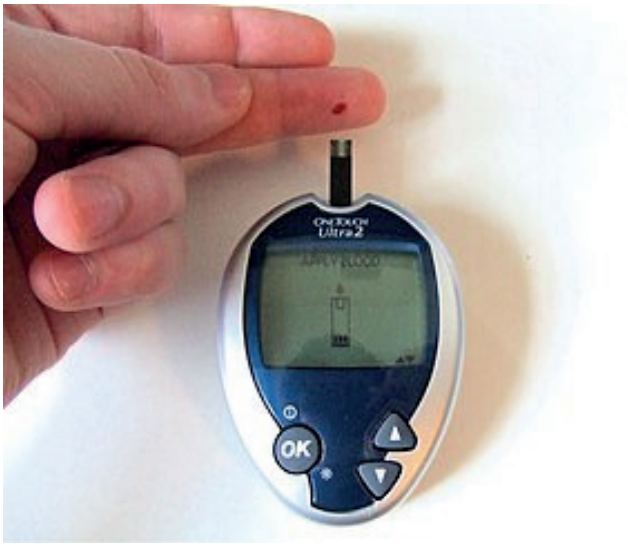
Does anyone have any final comments to add?

Thank you for your participation. Would everyone here feel comfortable being contacted in the future to be asked follow-up questions, or to help with this project further?

[Hand-out to participants]

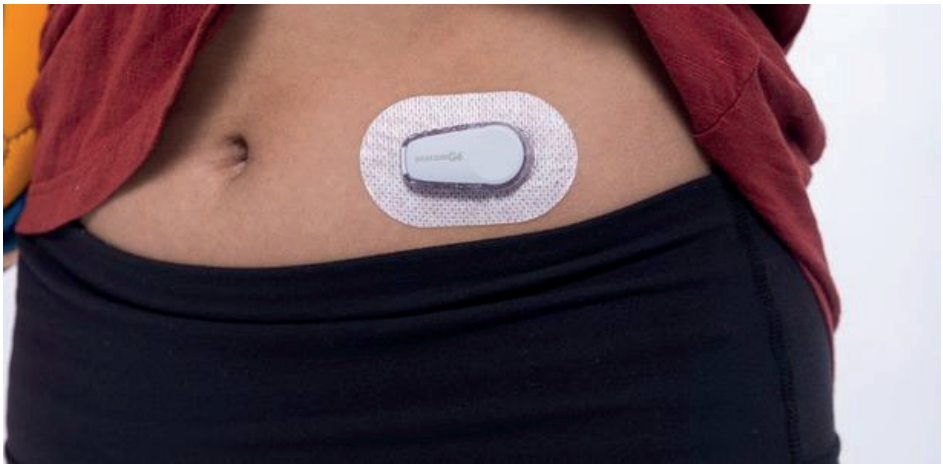
Finger-prick glucose monitors

(Finger-prick photo courtesy of TesaPhotography)

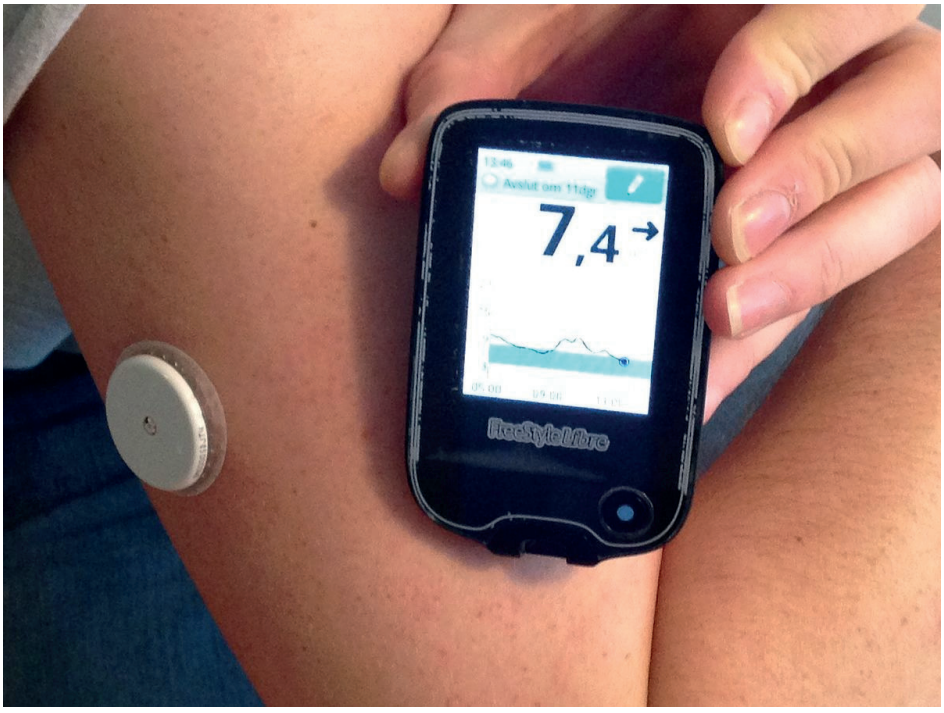


(Finger-prick blood glucose monitor photo courtesy of David-i98)

Continuous glucose monitors



(Dexcom G6 © Dexcom Deutschland GmbH)



(Flash glucose monitor photo courtesy of Sjö - Own work)

Focus group and Interview guidelines - FOR PATIENTS

Black = Speech to be directed to participants
[Italics] = Guidance for moderators or instructions
Numbers = Timing indications, how long each section should take

Instructions to moderators:

- *Create warm and friendly environment*
- *Interact with participants, and stimulate interaction between participants*
- *Make seating arrangements for participants according to their needs*
- *Exercise mild unobtrusive control (moderate the discussion but do not interrupt too often)*
- *Adequate knowledge of topic*
- *Have the discipline of listening and apply active listening*
- *Take into account the different types of participants and try to balance the conversation while addressing the obligatory topics: dominant talkers, shy participants, etc*

00.00 Welcome

Welcome, my name is…… and I will be your moderator guiding the discussion. Today we would like to discuss your thoughts about glucose monitoring devices, and any opinions, preferences, or concerns you might have. This discussion is part of a large European project called PREFER, which aims to make patients more involved in the development of their drugs or medical devices. The opinions collected today will be used to write reports and articles to inform companies, health authorities, and other researchers about what matters most to patients when it comes to their choices about glucose monitoring. Before we begin, I would like to ask if there are any questions you may have about the project or the purpose of the focus group today.

[Give participants the chance to ask for more information or any other questions they may have related to the focus group, the PREFER project, etc.]

If there are no more questions and everyone feels that they have enough information and would like to participate in the focus group, we will start by giving you a consent form to sign. Once this is done we will ask you to complete a short questionnaire about your background and diabetes experiences. Then we will ask you some questions. The whole focus group should take about an hour. We can have a short break after 45 minutes.

There are no right or wrong answers, we are looking for your personal opinions. It is possible that you might not agree with each other, but it would be nice if you could listen respectfully to each other. [*For Dutch study only* - We're audio recording, so we'd like to ask that only one person is speaking at a time. Also please put your mobile phones on silent. We will use first names today, but your names will not be used in any reports.] [*For Polish study only* - We will

create transcripts of what you type, but your names will not be used in any reports] If there are any questions or terms that are used during the focus groups that are not clear to you, please let us know.

[Give participants the consent form and questionnaire. Wait until everyone has handed back the documents before starting the next section.]

00.10

So, we will now start the focus group. Is that OK for everybody?
Let's go around the circle and say your first name and where you live.

00.12

Today we'll be talking about glucose monitoring devices. Is everyone familiar with what glucose monitoring is?

[Explanation if needed]

00.13 Is everyone familiar with why glucose monitoring is important?
[Explanation if needed]

00.14 There are lots of different kinds of devices that can help you monitor your glucose.

- Firstly, what is your experience with monitoring blood glucose?
- How do you feel about using a device to monitor your glucose?

00.20

There are several different kinds of glucose measuring devices, as you can see in your handout [*Indicate and point to picture]*. There are new devices becoming more available called continuous glucose monitors. These are small devices that have a tiny sensor that's inserted under the skin. This is attached to a transmitter that sends your blood sugar levels to a hand-held display device for you to look at, or even to your insulin pump if you have one. The device measures your glucose levels throughout the day and night, and lets you see trends over time, and can give you alerts if you are having high or low blood sugar. There is information and a picture in your participant information sheet

- What are your opinions about continuous glucose monitors?
- Would you consider using one?
- (*Follow-up:* Why or why not?)

- What do you think about when considering whether continuous glucose monitors are right for you?
- How do continuous glucose monitoring devices compare to devices that use a drop of blood from your finger?
- What do you think about when making a comparison?

00.25

Continuous glucose monitors would let you see your glucose levels over a long period of time. They can also help you tell you if your glucose is currently rising or dropping (and at what speed) helping you with timing your insulin, diet, and exercise

- How important is this to you?

00.30

These devices have a sensor that needs to go under the skin.

- How do you feel about that?
- How do you feel about how it looks?
- How do you feel about its size?

00.35

The sensor needs to be replaced every so often, depending on what kind of model and brand it is. Some sensors need to be replaced every two weeks, other every six months. With some models, you can insert the sensor into your skin by yourself, at home. With other models, you need a doctor.

- How would you feel about replacing the sensor yourself
- Would you prefer a doctor do it?
- How important is it that the sensor lasts a long time?
- What do you consider to be a good amount of time for a sensor to last?

00.45

Thank you everyone, we will now take a short 10 minute break. [*For Dutch study only:*] Please help yourself to tea and coffee

00.55

Continuous glucose monitors can give you an alert or alarm to warn you when your glucose is getting too high or too low

- How important is this to you?
- When might this be convenient?
- When might this be inconvenient?

01.05

Some continuous glucose monitors can be linked to an insulin pump, and automatically control your insulin depending on how high or low your blood glucose is

- How important would this be to you?

01.10

Some glucose monitors can be linked to your phone so that other people (family, partner) can know how you are doing?

- Is this something that appeals to you?
- How important would this be to you?

01.15

Some continuous glucose monitors still need you to check your blood glucose twice a day using a finger-prick test, to make sure it's measuring your glucose accurately. This is called calibration.

- Would this be inconvenient for you?
- There are other devices that don't need you to do this finger-prick check at all. But they are sometimes less accurate when your glucose is very low. Which seems more preferable to you?

01.20

Final questions now

- If you were having to choose a new glucose monitoring device, what is most important to you when making that choice?
- If you were having to choose a new glucose monitoring device, what is least important to you when making that choice?

01.30

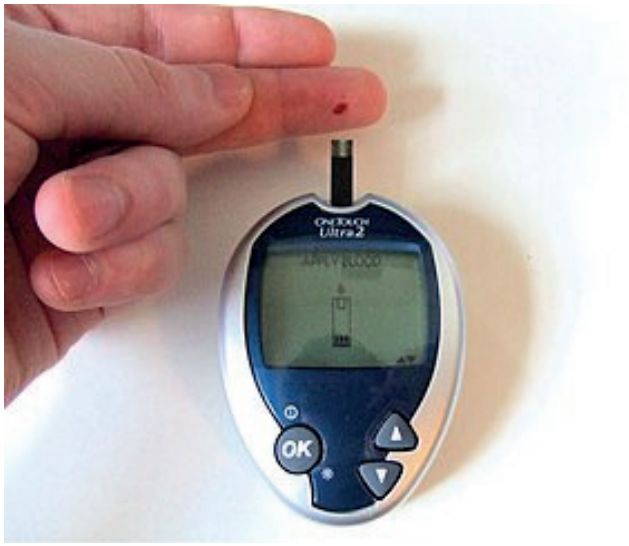
Does anyone have any final comments to add?

Thank you for your participation. Would everyone here feel comfortable being contacted in the future to be asked follow-up questions, or to help with this project further?

[Hand-out to participants]

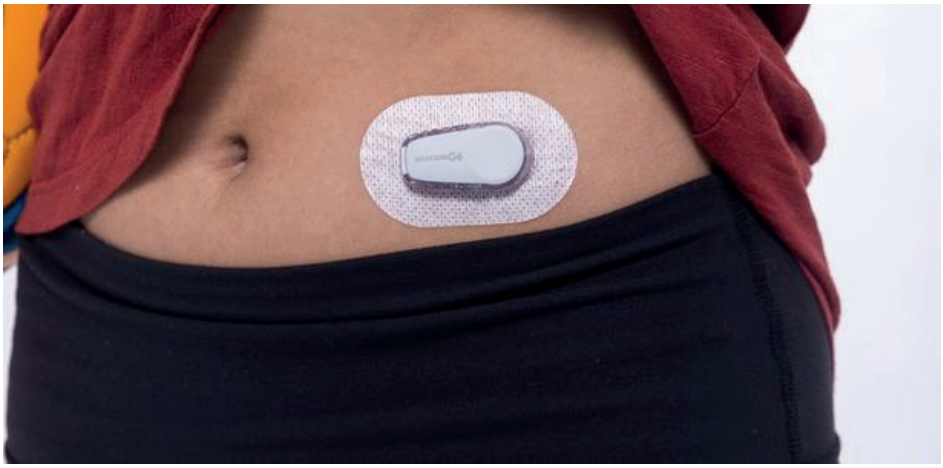
Finger-prick glucose monitors

(Finger-prick photo courtesy of TesaPhotography)

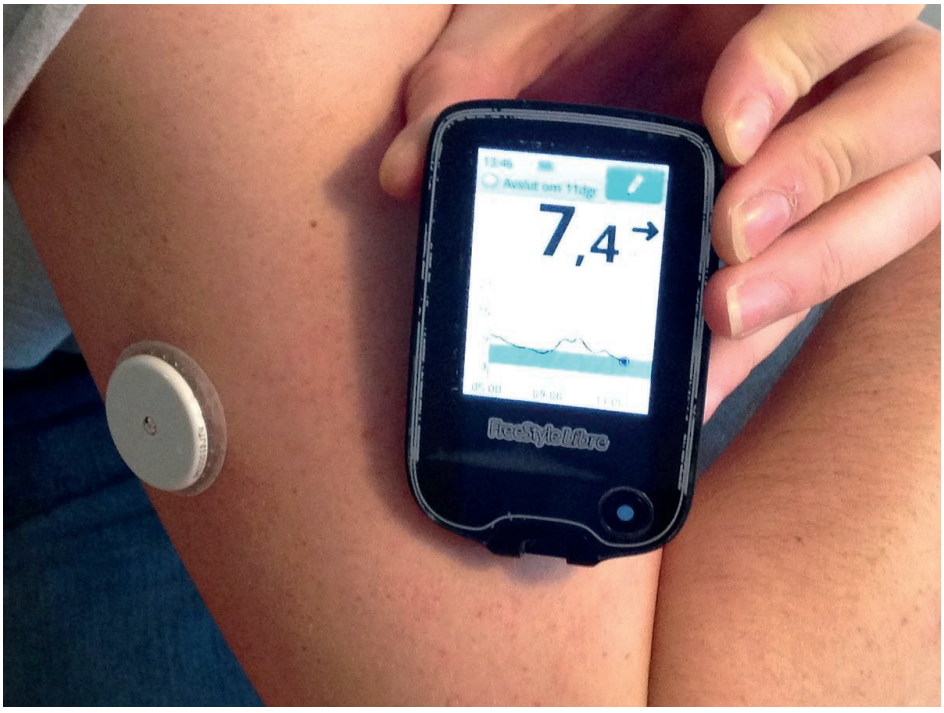


(Finger-prick blood glucose monitor photo courtesy of David-i98)

Continuous glucose monitors



(Dexcom G6 © Dexcom Deutschland GmbH)



(Flash glucose monitor photo courtesy of Sjö - Own work)

Chapter 4

A unique approach to identifying patient preferences for neuromuscular disorder treatments: a Q-methodology study

Authors: I.P. Smith, Cecilia Jimenez-Moreno, Cathy Anne Pinto, Rachel M. Ward, Emma-Jayne Ashley, Jane Newman, Alasdair Blaine, Chiara Whichello, Gráinne S Gorman, G.A. de Wit

ABSTRACT

BACKGROUND: Patient preference studies can inform medical product development, but often require large sample sizes, are cognitively burdensome, and may not identify less-common viewpoints. This makes it difficult to conduct preference studies within rare disease populations, especially in cases with impaired cognitive capacity. Q-methodology is a flexible preference elicitation technique that can overcome these issues.

OBJECTIVE: To give an example of the application of Q-methodology to identify shared opinions about unmet medical needs in patients with neuromuscular disorders.

METHODS: An online Q-methodology study was conducted in a sample of patients with myotonic dystrophy type 1 and mitochondrial myopathies, and their caregivers. Participants were recruited through patient organizations, cohorts, and registries in five English-speaking countries.

RESULTS: Seven factors representing clinically meaningful viewpoints about unmet medical needs were identified. Different groups focused on improving common symptoms and problems, less common symptoms and problems, or preventing future risks. 75.6% of patients and 90% of caregivers said it was 'easy' or 'very easy' to understand the Q-methodology questions.

CONCLUSIONS: Q-methodology represents a flexible method for assessing patient preferences in rare diseases. The use of this methodology can help ensure that patients and caregivers have a say in patient care.

INTRODUCTION

There has been growing interest in the use of patient preferences (PP) by industry and regulatory stakeholders in the development and evaluation of medical products [1, 2]. This type of information is generated by systematically asking patients what they value in patient care or in medical interventions, such as medicinal therapy [3]. This could be related to what benefits they want from treatment, associated costs, or acceptability of side-effects. [2, 4-7]. PP can be applied to new medical products across all phases of the medical product lifecycle, from research into new potential medicines, improving options currently available, or allowing products onto the market for patient care [8, 9].

Commonly used techniques to generate PP, like discrete choice experiments (DCE), often rely on large sample sizes [10, 11], can potentially be cognitively burdensome or difficult to administer [12, 13], and are not always suitable for patient populations with small but specific sub-groups who may have different care needs [14, 15]. The lack of suitable patient preference elicitation methods has led to scenarios in which decision-making for new medicines or evaluation of potential treatments for rare diseases has been based on limited patient information [16].

One possible technique which could help to overcome these issues is Q-methodology. Q-methodology is a combined qualitative and quantitative technique which can be used to empirically study distinguishing viewpoints in patient populations [17-19]. Q-methodology aims to capture the subjectivity of participants' preferences while identifying correlations across the participant sample, allowing researchers to contrast different attitudes or perspectives. While this technique was originally developed nearly a century ago, it has only been recently applied to the field of healthcare in response to a recognition that a better understanding is needed not only of what perspectives stakeholders have, but also why they have them [20]. Three benefits of Q-methodology studies are that a) the findings are robust even with small patient groups, b) small groups of patients within the population are identifiable[14], and c) Q-methodology tasks are relatively simple in nature with the complexity easily adjustable to target population needs.

Traditionally, Q-methodology studies have been conducted in face-to-face interviews with a researcher present [20]. Recently, online versions of Q-methodology have become more common [21], but there is limited evidence supporting the use of online Q-methodology studies in patient populations. The aim of this study is to identify common opinions about unmet medical needs in a sample of patients with rare diseases and their caregivers using Q-methodology.

MATERIALS AND METHODS

Case Study

This case study assessed the unmet medical needs of patients with neuromuscular disorders (NMDs), i.e. myotonic dystrophy type 1 (DM1) and mitochondrial myopathies (MM) in an online survey was used for this study [22]. These NMDs are generally characterized by impaired muscle functioning [23, 24], but the individual symptomologies can be diverse and include other complications such as limited physical functioning, fatigue, and cognitive impairment [23-29]. Little work identifying the preferences of these diseases has been done due to how rare these diseases are (approximately 5 to 20 individuals per 100,000 [23, 30, 31]), the large number of different clinical phenotypes, and the associated cognitive impairment, which make it difficult to assess preferences using common preference elicitation techniques [16].

Participants were recruited through social media and website advertisements posted by NMD patient organizations, through email or post invitations sent by patient registries in five English-speaking countries (UK, US, Canada, New Zealand, and Australia) [22]. To be included in this study, all participants had to self-report a diagnosis of DM1 or MM with an early onset of disease (<20 years of age) or be the caregiver of a patient matching that description, be able to give informed consent to participate, be able to read and understand English, and be able to complete an online questionnaire. Participants were excluded if they reported a historical diagnosis of encephalopathy or dementia.

The study was conducted in accordance with the UK Policy Framework for Health and Social Care Research and the new EU General Data Protection Regulation (GDPR) and was approved by the Newcastle University Ethics Committee (Ref: 15169/2018) and The North West – Greater Manchester West Research Ethics Committee (Ref: 20/NW/0367). All participants provided informed consent prior to participation in the study. As an incentive, patients could donate a small amount of money to a patient organization of their choice or receive a gift card.

Q-Methodology

In a Q-methodology study, participants are given a set of statements (the Q-set) on a specific topic. Participants are asked to sort these statements on a relevant ranking dimension, such as how important or unimportant the statements are to the participant, whether they agree with the statement, and whether the participant relates to the statement [17, 19]. How participants rank the statements in relation to each other is called a 'Q-sort'. After the sorting exercise, participants are asked to reflect on why they decided to sort the statements in the way that they did in order to better understand the reasoning behind their perspective. This is usually done in in a qualitative interview or open-field text box. The resulting Q-sorts are then analyzed using a by-person factor analysis which identifies commonalities between people rather than between instrument variables, such as done in more traditional by-variable factor analysis [20]. The

resulting factors represent shared opinions between the participants and are interpreted using the average statement rankings for each factor [32, 33]. This interpretation, combined with qualitative information about why a participant ranked the statements in such a way, provides a deeper understanding of not only what shared opinions exist, but also why participants have this opinion.

Development of the statement set

The statements used in the survey were developed in a qualitative study assessing unmet healthcare needs of DM1 and MM patients and caregivers as well as potential side effects of hypothetical treatments [34]. In this qualitative study, 52 individuals were interviewed (15 individual interviews and 5 focus groups with 5 to 9 participants each) including 33 patients and 19 caregivers, representing both disease groups. In these interviews and focus groups, unmet healthcare needs of patients were identified based on disease signs and symptoms, their impact on daily life activities, and potential side effects of hypothetical treatments. Additionally, caregivers were asked about the medical needs of patients who could not self-report due to cognitive limitations to ensure that these needs did not go unrecognized. This resulted in an initial list of 11 unmet healthcare needs and 6 potential risks of new, hypothetical treatments. To maintain mutual relevance, the list was reduced by excluding any unmet need specific to either DM1 or MM patients (e.g. myotonia sign or hearing loss). The final list consisted of 9 symptom targets for improvement inn treatment and 2 potential risks of side-effects of treatment that DM1 and MM patients and caregivers were concerned about (Table 1).

Table 1. Statement Rankings by Factor

	Factor						
	1	**2**	**3$^{\beta}$**	**4$^{\beta}$**	**5**	**6**	**7**
1. Improved muscle strength*	+2	0	+1	0	0	0	+2
2. Improved energy and endurance ***	+2	0	+2	**-1**	+1	+1	+1
3. Improved balance	**0**	-2	-2	-1	**+2**	-2	-1
4. Improved cognition	-2	**+1**	0	0	0	0	-2
5. Improved speech and communication	-1	+1	**-2**	**+2**	+1	+1	0
6. Improved gut function	0	-1	-1	-2	-2	**+2**	0
7. Improved pain in joints or muscles	0	-1	+2	+1	+2	+2	0
8. Improved swallowing of liquids and food	**+1**	0	-1	**+2**	-2	-1	-2
9. Improved heart and cardiovascular health**	+1	+2	0	0	0	0	+2
10. Lower risk of temporary blurry vision$^{\mathbf{T}}$	-2	-2	0	+1	-1	-2	+1
11. Lower risk of permanent liver damage	-1	**+2**	**+1**	-2	-1	-1	-1

Higher or lower loadings indicate greater or lower importance (respectively) for this factor, 0 indicates that this statement was neutral for the participants loading onto this factor; Defining statements are shown in bold; β: Bipolar factors indicating that Q-sorts loaded as either strongly agreYeing with or strongly disagreeing with this factor; * ranked as most important by 38.0% of participants, ** ranked as most important by 29.6% of participants, ***ranked as most important by 26.8% of participants; **T** ranked as least important by 54.9% of participants

Q-methodology task

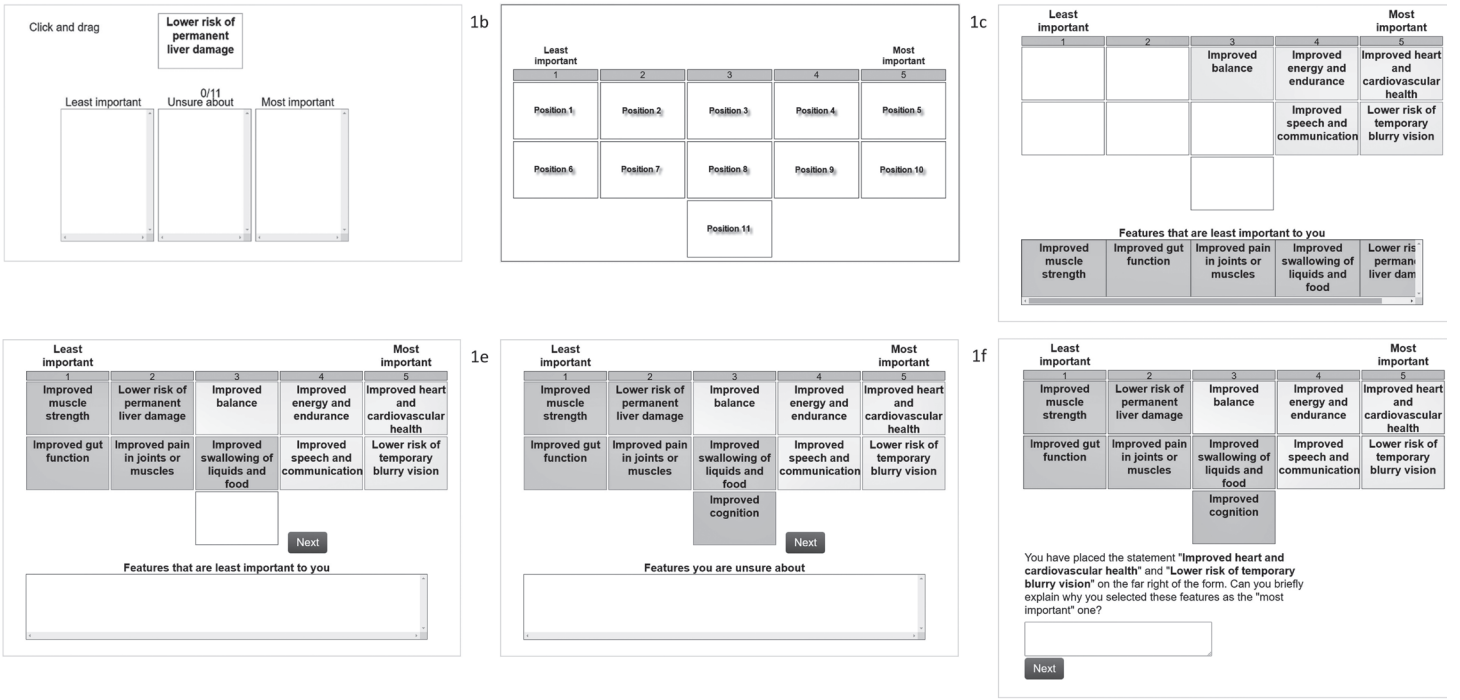
In our study, participants were asked to imagine they were talking with their doctor about hypothetical new treatments for their disease and the associated benefits and risks or to consider these benefits and risks from the perspective of the patient they care for. The participants were instructed that they would be shown statements which they would have to sort into groups. The sorting process consisted of two steps. First, participants sorted the statements according to three broad categories of most important, least important, or unsure about (Figure 1A). This initial task familiarized the participant with the statements and simplified the second more detailed sorting task. In the second task, participants were asked to sort the comments from each category onto a five-column factor array with the column on the left-hand side indicating the least important statements and the right-hand side column indicating most important statements (Figure 1B). Participants sorted all the statements until the factor array was full (example Figure 1C-E). After the sorting was finalized, the participants were asked to describe what guided their opinion and sorting for their most and least important statements (Figure 1F).

Q-methodology analysis

Q-sorts were analyzed using an iterative by-person principal-factor analysis with an Oblimin rotation, to identify participants with similar opinions about what is important in potential treatments [17]. The Oblimin rotation allows for correlation of the factors which were as expected due to the limited number of statements used in the survey and overlapping symptomologies [35]. Statements "loaded" either positively (+1 = somewhat important; +2 = very important), negatively (-1 = less important; -2 = least important), or neutrally (0) onto each factor, based on the common rankings expressed. "Defining statements" were identified if they loaded onto a factor in a way that was unique from other factors, such as when a statement was only important for one group. Due to the limited sample size, factors were allowed to be bipolar, reflecting an opinion which a patient can either agree with or be opposes to.

The optimal number of factors extracted from the analysis was chosen using the following criteria: 1) all factors were required to have an eigenvalue > 1.00 [36, 37]; 2) all positive factor correlations were $d < 0.50$ to prevent highly similar factors but not highly dissimilar factors [38]; 3) each factor had a minimum of two Q-sorts that significantly loaded only to this factor (i.e. at least two participants only associated with the factor; and 4) the factors needed to represent a clear, distinct, and theoretically meaningful common viewpoints about what is important in hypothetical treatments based on a final qualitative assessment by clinical healthcare providers (N=3) and patient experts (N=2).

Figure 1. Example of a sorting task



1a Initial sorting in broad categories; 1b Empty grid before sorting statements; 1c Grid with most important statements sorted according to importance; 1d Grid with least important statements sorted according to importance; 1e Grid with all statements sorted according to importance; 1f Example of post sorting qualitative process to explain why statements were sorted as being "most important"

Weighted Q-sort factor arrays were created by averaging the participant Q-sorts which were significantly associated with the factor [17]. These factor arrays can be interpreted as shared opinions about unmet medical needs for DM and MM patients and represent the opinion that a person would have if they had perfectly loaded onto a factor. Qualitative information from individual Q-sorts associated with the factors was used to better understand why these opinions were held. A p-value of 0.05 was used as a cutoff for significance. The interpretation of the factors was supported by including the statement number and their importance to the factor, where reported. The principle-factor analysis was done using the qfactor package in Stata 14.2 [39].

Additional Survey Information

The survey started by collecting information related to patient demographics and clinical history questions. This was followed by a short animated educational video with a voiceover narration of the text information (~5 minutes in length) to instruct participants on how patient preference research is used and how to complete the Q-methodology task. The participants then completed the Q-methodology task. Finally, participants were asked to rate how easy or hard it was to understand and complete the task, and whether the instructional video was helpful or not.

RESULTS

Study sample

In total, N=72 completed Q-methodology tasks were received. One person reported having not understood the exercise and filled in the array incorrectly. Their response was removed from the final analysis resulting in a final set of N=71 Q-sorts received from patients (N=14) or caregivers of patients (N=16) with DM1, and patients (N=27) or caregivers of patients (N=14) with MM. Of the caregivers, N=6 of the 14 caregivers reported having themselves a MM or DM1 diagnosis or suffering from disease-related symptoms themselves. The demographics of the participants by disease type and role can be found in Table 2.

Statement Rankings and Factors

Seven unique and interpretable factors were identified after consulting with clinical and patient experts. These factors accounted for 62.2% of the variance (Table 1). Of the 71 responses, N=39 (54.9%) were

Table 2. Demographics of participants by Disease Type and Role (Patient or Caregiver)

	Myotonic Dystrophy Type 1	Mitochondrial Myopathies
Patient or Caregiver (N, %)*	14 (46.7)	27 (65.9)
Patient	16 (53.3)	14 (34.2)
Caregiver		
Patients	(N=14)	(N=27)
Age: M (SD, range)	42.1 (14.0, 20-67)	47.8 (17.2, 20-77)
Gender (% Female)	71.4	55.6
Country of origin (N, %)		
United Kingdom	5 (35.7)	15 (55.6)
United States	6 (42.9)	9 (33.3)
Canada	2 (14.3)	1 (3.7)
New Zealand	1 (7.1)	0 (0.0)
Australia	0 (0.0)	0 (0.0)
Other	0 (0.0)	2 (7.4)
Work Status (N, %)**	5 (35.7)	4 (14.8)
Employed Full-time	3 (21.4)	1 (3.7)
Employed Part-time	1 (7.1)	2 (7.4)
Do voluntary work	2 (14.3)	9 (33.3)
Do not work or volunteer due to my disease	0 (0.0)	2 (7.4)
Unemployed for other reasons	1 (7.1)	3 (11.1)
Student	2 (14.3)	6 (22.2)
Retired		
Educational level (N, %)		
Attended a school for special education	0 (0.0)	0 (0.0)
Attended formal schooling up to age 14	1 (7.1)	2 (7.4)
High school graduate, diploma or the eq	2 (14.3)	4 (14.8)
Some college credit or trade/technical/	3 (21.4)	9 (33.3)
Bachelor's degree	3 (21.4)	10 (37.0)
Professional degree	1 (7.1)	0 (0.0)
Master's degree	3 (21.4)	1 (3.7)
Doctorate degree	1 (7.1)	1 (3.7)
Motor function (N, %)		6 (22.2)
can walk AND run without the need of an assistive device	6 (42.9)	8 (29.6)
can walk without the need of an assistive device	3 (21.4)	4 (14.8)
can walk but rely on an assistive device	3 (21.4)	4 (14.8)
can walk (aided or unaided) but use a wheelchair part-time	2 (14.3)	5 (18.5)
rely fully on a wheelchair	0 (0.0)	
Caregivers	(N=16)	(N=14)
Age: M (SD, range)	46.6 (13.3, 27-71)	47.9 (10.8, 35-67)
Gender (% Female)	68.8	85.7
Gender of patient (% Female) ***	30.0	44.4
Country of origin (N, %)		
United Kingdom	5 (31.3)	9 (64.3)
United States	8 (50.0)	4 (28.6)
Canada	1 (6.3)	0 (0.0)
New Zealand	1 (6.3)	0 (0.0)
Australia	0 (0.0)	1 (7.1)
Other	1 (6.3)	0 (0.0)

(cont'd)

	Myotonic Dystrophy Type 1	Mitochondrial Myopathies
Work Status (N, %)		
Employed Full-time	7 (43.8)	7 (50.0)
Employed Part-time	3 (18.8)	2 (14.3)
Do voluntary work	0 (0.0	0 (0.0)
Do not work nor volunteer due to my disease	2 (12.5)	2 (14.3)
Unemployed for other reasons	1 (6.3)	1 (7.1)
Student	0 (0.0)	0 (0.0)
Retired	3 (18.8)	2 (14.3)
Education level (N, %)		
Attended a school for special education	0 (0.0)	1 (7.1)
Attended formal schooling up to age 14	1 (6.3)	0 (0.0)
High school graduate, diploma or the eq	3 (18.8)	3 (21.4)
Some college credit or trade/technical/	4 (25.0)	2 (14.3)
Bachelor's degree	3 (18.8)	3 (21.4)
Professional degree	1 (6.3)	3 (21.4)
Master's degree	3 (18.8)	1 (7.1)
Doctorate degree	1 (6.3)	1 (7.1)
How long have you been a caregiver? (N, %)		
≤5 years	2 (12.5)	5 (35.7)
6-10 years	1 (6.3)	1 (7.1)
11-20 years	4 (25.0)	1 (7.1)
>21 years	3 (18.8)	2 (14.2)
No answer	6 (37.5)	5 (35.7)
Relationship with the patient? (N, %) *	14 (82.4)	14 (100.0)
Parent	2 (12.5)	0 (0.0)
Other close relative or adult living in the same house		
Gender of patient (% Female) **	30.0	44.4
Age of onset (N, %)		
Since birth	3 (18.8)	2 (14.3)
0-5 years	4 (25.0)	4 (28.6)
6-10 years	2 (12.5)	0 (0.0)
11-20 years	1 (6.3)	3 (21.4)
21-30 years	0 (0.0)	0 (0.0)
>30 years	0 (0.0)	0 (0.0)
No answer	6 (37.5)	5 (35.7)
Motor function of patient cared for (N, %)		
can walk AND run without the need of an assistive device	4 (25.0)	3 (21.4)
can walk without the need of an assistive device	1 (6.3)	0 (0.0)
can walk but rely on an assistive device	0 (0.0)	0 (0.0)
can walk (aided or unaided) but use a wheelchair part-time	1 (6.3)	1 (7.1)
rely fully on a wheelchair	0 (0.0)	1 (7.1)
No answer	10 (62.5)	9 (64.3)

*N=6 caregiver participants reported also being diagnosed with myotonic dystrophy (N=3) or another type of mitochondrial disorder (N=3); ** *Unemployed for other reasons* includes N=3 participants who responded that they were full-time caregivers, *Student* includes N=1 participant that follows a study part-time, *Retired* includes those that reported being medically retired; *** % of responses given; uniquely associated with only one factor with six of the seven factors having a defining statement. The largest factor (Factor 1) was associated with N=12 participants and the smallest factors (Factors 6, 7) with N=3 participants. Two of the factors (Factor 3 and Factor 4) were bipolar in their loading, indicating that they represent two opposing opinions. No statements were found in the analysis with a consensus opinion about importance. No single factor was associated with only patients or caregivers, but factor 7 was only associated with MM patients or caregivers (Table 3). Three of the 7 factors were associated with participants who rated risks of treatment as least important compared to the potential benefits of

Table 3. Role and disease classifications for each factor

	No Loading	Factor							Total Loaded	Total
		1	**2**	**3**	**4**	**5**	**6**	**7**		
All Patients n (%)	17 (41.5)	7 (29.2)	4 (16.7)	3 (12.5)	4 (16.7)	3 (12.5)	1 (4.2)	2 (8.3)	24 (58.5)	41
All Caregivers n (%)	15 (50.0)	5 (33.3)	4 (26.7)	1 (6.7)	1 (6.7)	1 (6.7)	2 (13.3)	1 (6.7)	15 (50.0)	30
DM1 Patients and Caregivers n (%)	12 (40.0)	6 (33.3)	7 (38.9)	1 (5.6)	2 (11.1)	1 (5.6)	1 (5.6)	0 (0.0)	18 (60.0)	30
MM Patients and Caregivers n (%)	20 (48.8)	6 (28.6)	1 (4.8)	3 (14.3)	3 (14.3)	3 (14.3)	2 (9.5)	3 (14.3)	21 (51.2)	41
Total	32 (45.1)	12 (30.7)	8 (20.5)	4 (10.3)	5 (12.8)	4 (10.3)	3 (7.7)	3 (7.7)	39 (54.9)	71

DM1: Myotonic Dystrophy Type 1; MM: Mitochondrial Myopathies

a treatment. The remaining factors were only concerned with reducing one potential risk of a treatment, of which only Factor 2 thought that risk reduction was a high priority.

On average, Statements #2 (improved muscle strength), #9 (improved heart and cardiovascular health), and #1 (improved energy and endurance) were ranked as the most important unmet medical need, and were ranked as most important 38.0%, 29.6%, and 26.8% of the time, respectively. Statement #10 (lower risk of temporary blurry vision) received the lowest ranking of importance on average and was listed as least important 54.9% of the time. All statements were ranked at every level of importance by at least one participant except for lower risk of permanent liver damage which was never ranked positively by any caregiver or patient living with DM1. Improved heart and cardiovascular health (#9) never loaded negatively in any factors, whilst improved cognition (#4) mostly loaded negatively or neutral. Both were neutrally loaded (neither important nor less important) for four of the seven factors (3, 4, 5, 6).

Factor Descriptions

Factors are described using the statement numbers (#) and how they loaded on to the factor, ranging from +2 to 0 to -2 (for most important, neutral, or least important respectively). Differences in the factors were often related to the symptoms that the patients or the caregivers were currently experiencing, although there were some instances where future concerns were prioritized. Example statements explaining why participants ranked the statements as most or least important for each factor can be found in table 4. The largest factor (Factor 1, n=12) was evenly split for disease classification and had the most patients and caregivers associated with it. This group tended to focus on aspects of treatments that related to strength or energy and endurance rather than cognitive or long-term issues. For this group, physical capability was the most important area of focus (#1, +2; #2, +2; #9, +1), but they were also concerned with simple activities of daily living, such as being able to swallow liquids and food (#8, +1). However, there was no concern with cognitive functioning (#4, -2), and they found risks associated with possible treatments less important than potential benefits (#10, -2; #11, -1). Uniquely, this group was neutral on improving balance where most other groups saw this as either not important, or the aspect of highest importance.

The second largest factor (Factor 2, n=8) was evenly split between patients and caregivers but primarily consisted of participants from the DM1 group. This group tended to focus on how the treatments would impact organ functioning (as this would fundamentally affect their health with cardiovascular health)(#9, +2) with a primary concern of lowering risk of liver damage (#11, +2) being their primary concerns. In line with this focus on physiological functioning over physical abilities, this group was also unique in prioritizing the improvement of cognitive functioning (#4, +1) and speech and communication (#5, +1).

The remaining factors were associated with fewer participants (N=3-5 per factor) but represented unique opinions about what is important regarding unmet medical needs. These factors concentrated on muscle symptom concerns (Factor 3), speech and swallowing difficulties (Factor 4), movement and mobility concerns (Factor 5), basic functioning and quality of life (Factor 6), and attention on underlying health to support functioning associated with MM patients only (Factor 7). Distinguishing aspects of these factors included a focus on pain (guiding issue for Factor 3), a lack of relative importance for improving energy and endurance or muscle strength (Factor 4), a focus on balance as a reason for improving pain in joints (Factor 5), and the impact of gut functioning on quality of life (Factor 6).

Patient assessment of task

Table 5 shows the patient feedback to the Q-task. Participants who answered these questions generally found the survey easy to understand (81.7%) and to answer (81.7%) regardless of whether they were a patient or caregiver, or which disease group they identified with (table 5). Only four participants (5.6%) said it was difficult or very difficult to understand the questions, three of which were caregivers. Similarly, six participants (8.4%) responded that it was difficult or very difficult to answer the questions, four of which were caregivers. Most participants said the instruction video helped them to understand the Q-methodology questions to some degree (69.0%).

DISCUSSION

This study focused on the use of Q-methodology to identify unmet medical need prioritization of patients with DM1 of MM and their caregivers using an online questionnaire format. The Q-methodology task identified seven clear, distinct, and theoretically meaningful common viewpoints according to clinical and patient experts. The statements most reported as important tended to match that of the most common symptoms of both DM1 and MM, such as reducing pain and improving physical strength and endurance [23, 24, 40-47]. Most of the participants, regardless of patient or caregiver status or disease group found the Q-methodology easy to understand and answer and found the informational video helpful. The online setting was well accepted by most of the participants and enabled us to reach a much larger, dispersed patient population in five different countries.

Table 4. Examples of reasons for why statements were ranked as most and least important

Factor	Supporting statements for statement ranking (copied directly from surveys)	
	Most Important	**Least Important**
1	"Energy levels are low and I would like to live more normally"; "I have trouble doing some day-to-day activities due to weakness"; "I want to be able to take care of myself. I choke on food and drink. It"s very scary."; "So I can do more , So i can eat properly without a peg"; "These are the two most important factors that have the most effect on my way of life and the inability to do activities I use to enjoy or would like to do now."	"My cognition is fine"; "These issues are important, but they are less important than overall health and wellness"; "Temporary blurry vision is okay since it is not permanent. Although permanent liver damage is not something I"d like to risk, I had to put it in order of importance and could not put it into a different category, there was no option."; "Hopefully with improved muscle strength there would also be improved balance."
2	"Cardiac health can severely impact how I am able to live"; "Permanent liver damage is not an acceptable risk for a myotonic dystrophy drug unless the drug was a near cure"; "Improving the cognition would make life easier and increase his QOL"; "Heart , lungs, liver all vital organs"	"I can live with temporary blurry vision and improved balance is important but not as important of the other things"; "The person I care for already has these issues and while they are a problem, they are minor in the grand scheme of things and are already dealt with as part of his day-to-day. While an annoyance they are not as debilitating as other issues."
3[b]	"My two worst symptoms are overwhelming fatigue and severe muscle pain, so it is most important to me that I would experience improvement in these two areas."; "These are the ones that cause the most quality of life issues";	"My balance problems don"t have a big impact on my quality of life"; "His lifestyle does not require much energy or endurance."
4[b]	"I feel as though the item is lodged in my throat and I can"t breathe"; "I have too many friends with mitochondrial disease that have died from sepsis! "; "These two problems can severely compromise life expectancy whereas loss of balance etc. can be lived with"; "My gut function is the symptom I have that"s most likely to be life-threatening, if I eventually need to go on TPN"	"If I have to pick for my life now, I would prefer to maintain my cognition and gut function over muscle strength."; "My swallowing is an issue but not a very big one"; "My daughter doesn"t really have these symptoms right now so they are not as important to me personally as the others. However, if she started to have issue swallowing or chronic pain, I"m sure those would move to the top of my list fast."
5	"If he were able to communicate and do things for himself, (he and we) would be really happy."; "These are the areas most affected"; "movement is very important to me"; "Because my main symptom of Mito is Ataxia and Mycolnas. These areas are the most affected."	"Additional medication can be taken to regulate gut function."; "My son does not suffer from either of those things"
6	"Pain from dystonia is a major feature of his condition so less pain would be so important. He has severe reflux and vomiting at times so better gut function would be a big help.; "These symptoms significantly impact daily life and improvements would improve quality of life."	"His muscles are actually extremely strong because of dystonia so improving muscle strength is not needed, although better active coordinated voluntary movement would be."; "Patient is PEG fed so can do without swallowing liquids and foods."; "Temporary blurry vision while not ideal would not significantly impact daily life as patient is under full care"
7	"We believe if the heart and cardiovascular function works well the mitochondria will not place as much of a strain upon the heart keeping it as healthy as possible is a primary objective."; "These are the problems that I experience the most and find most difficult"	"We do not experience real issues either with cognition or swallowing liquids or foods"; "Liver damage is my least important as it isn"t a problem for me currently"

b: Bipolar factors indicating that Q-sorts loaded as either strongly agreeing with or strongly disagreeing with this factor

Table 5. Participant Feedback on Ease of Understanding and Answering the questions and instruction video

	How easy or difficult was it for you to…						Did the instruction video help you understand how to do the questionnaire?		
	Understand the questions?			**Answer the questions?**					
	Easy or Very Easy	Not Easy or Difficult	Difficult or Very Difficult	Easy or Very Easy	Not Easy or Difficult	Difficult or Very Difficult	Very Much	Moderately or a Little	Not at all
All Patients	75.6	22.0	2.4	82.9	9.8	7.3	51.2	36.6	9.8
All Caregivers	90.0	0.0	10.0	80.0	10.0	10.0	36.7	60.0	3.3
DM1 Patients and Caregivers	90.0	6.7	3.3	86.7	6.7	6.6	36.7	56.7	6.7
MM Patients and Caregivers	75.6	17.1	7.3	78.1	12.2	9.7	51.2	39.0	7.3

DM1: Myotonic Dystrophy Type 1; MM: Mitochondrial Myopathies

To our knowledge, this is the first study to use an online Q-methodology study in a patient group that may suffer from cognitive difficulties, which adds support to the use of online Q-methodology as a preference task for assessing patient priorities in rare disease groups [20, 21]. Participants generally reported that improvements in symptoms were a greater concern to them than potential associated risks of treatments, although avoiding risks were still important to many participants. While the largest factors reflected the most common symptoms of both DM1 and MM as the most important unmet medical needs, the strength of Q-methodology can be found in Factors 4 through 7. These factors found small but unique clusters of 3 to 5 participants related to less common phenotypes with unique priorities for care. For example, Factor 4 was the only factor in which swallowing problems and improved speech and communication were prioritized by patients. While not the most recognized symptoms of NMD, swallowing issues are not uncommon with studies reporting difficulties in upwards of 35% of patients and some reports have over 50% of patients reporting issues in communicating [45-47]. Some participants in this study shared that symptoms associated with swallowing difficulties were “scary” and made them feel like they “can’t breathe”. This distinct group of patients would likely not have been identified using more common preference elicitation methods. Ignoring these priorities when developing new treatments may result in a significant proportion of the patient population suffering while more commonly recognized symptoms are focused on. Interestingly, there were numerous instances when participants considered issues they may experience in the future and prioritized these based on a comparison of current symptoms and possible disease progression. This was not altogether unexpected but highlights

the ability of Q-methodology to not only identify the impact of current symptoms, but also the potential concerns that patients and caregivers may have for the future.

It should be noted that the aim of a Q-methodology is exploratory rather than comprehensive population assessment, thus the results are not meant to represent the priorities of the entire population [48]. This is partially reflected in how 45.1% of participants did not load onto only one factor. However, while there were several participants who did not load uniquely onto one factor the high amount of variance explained by the factor analysis indicates that this is likely the result of participants loading onto multiple factors rather than missing a distinct shared opinion. Furthermore, participants had to rank some aspects as less important even if they found them to be important due to the forced nature of the ranking method. Outcomes should thus be interpreted in regard to the relative importance of the different aspects rather than as an independent valuation of each aspect itself.

This study did have some weaknesses that are relevant in interpreting the outcomes. First, the set of statements was smaller than what is traditionally used in Q-methodology studies (typically around 40)[20, 49]. Although Q methodology can accommodate larger sets of statements, the number of statements for this study was intentionally kept small to account for the anticipated cognitive burden for patients that a larger set of statements would present in an online setting without an interviewer present [50, 51]. Further research may examine whether NMD patients are able to complete Q-methodologies with larger sets. The limited number of statements means that this study may lack some of the depth of insight that a more traditional Q-methodology study gives. Second, while the total sample size was comparable to other Q-method studies, the sample was split between patients and caregivers, and between disease groups. Ideally, more participants would have been included from each individual group (DM patients, DM caregivers, MM patients, MM caregivers) to support the unique opinions found. Fourth, Q-methodology is traditionally done face-to-face which allows for interaction between interviewer and participant. Listening to the participants while they complete the Q-sort offers additional insights to the individual aspects influencing their sorting decisions which we were not able to collect. Further, face-to-face interviews may also allow for the inclusion of those with more congenital or more severe DM and MM who may not have been included in the current survey. These patients may not be able to complete an online survey on their own due to greater difficulty with cognitive tasks.

Future research could expand on the use of Q-methodology with more statement sets applied in larger samples, to see if this results in more nuanced factors. For instance, to explore whether there are specific aspects to improving muscle strength that patients are more concerned about, such as wanting to do more household tasks or to not be perceived as incapable. This deeper understanding of NMD symptoms, and the impact on the wellbeing of patients, may (in the absence of pharmaceutical based treatment) help with the implementation of other support and care services, ultimately improving quality of life. It should be emphasized that while a

high number of patients found the method easy or very easy to complete, care should be taken to ensure that more complex statement sets are understandable and well tested. Additionally, while most participants reported that the task was easy to understand and complete, there was still a small group of participants who said that they had difficulties, including one participant who stated they did not understand the task. Future studies should consider an online guided interview setting where the Q-methodology task is completed with a researcher present. This flexible research design would allow for both the benefits of the interviewer led traditional design with the benefits of the online design, namely reach and ease of data collection. The growing ubiquity of mobile devices with video-calling capabilities across all sociodemographic strata should allow for this type of hybrid design.

CONCLUSION

Q-methodology was effective in identifying unique opinions and concerns regarding unmet medical needs for patients with NMD and their caregivers. The fully online survey was able to reach a geographically diverse sample of patients that would not have been possible using traditional face-to-face designs. Future Q-methodology studies should look at making the statement sets larger and more nuanced and use a hybrid design in which interviewers use video-calling technologies enabling them to be present during the tasks. This will result in richer qualitative information while ensuring that those patients with cognitive impairments have more opportunity to participate in research, and have their priorities identified. Q-methodology is a flexible tool that can help researchers identify and understand preferences from patients with rare diseases, helping to ensure that all patients regardless of location have their views and needs acknowledged, and can have a say in their care.

ACKNOWLEDGEMENTS

The authors of this manuscript would like to acknowledge the following people for their contributions to the study: We would like to thank all patient organizations supporting this study: MDUK, MDSG; Cure DM CIC; the Lily Foundation for mitochondrial disorders; UMDF, MitoCanada; MDC; MDA; MDF; and, MDNZ. We thank members of the Wellcome Research Centre for Mitochondrial Research that provided input and expertise into the design of this study. We thank the UK Myotonic Dystrophy Patient Registry at Newcastle University for facilitating participants recruitment and supporting with the pilot stages of the study and the Newcastle Wellcome Centre for Mitochondrial Research Patient Cohort for their assistance in reaching out to patients and caregivers (REC Ref: 13/NE/0326). Eline van Overbeeke for her assistance during the qualitative interview stage of the study.

REFERENCES

1. de Bekker-Grob, E.W., et al., Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public-Private Project. Patient, 2017. **10**(3): p. 263-266.
2. Levitan, B., et al., The Ball is in Your Court: Agenda for Research to Advance the Science of Patient Preferences in the Regulatory Review of Medical Devices in the United States. Patient, 2017. **10**(5): p. 531-536.
3. Administration, U.D.o.H.a.H.R.F.a.D., Patient-Focused Drug Development: Collecting Comprehensive and Representative Input; Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders, Center for Drug Evaluation and Research, Editor. 2020.
4. Ho, M., et al., A Framework for Incorporating Patient Preferences Regarding Benefits and Risks into Regulatory Assessment of Medical Technologies. Value Health, 2016. **19**(6): p. 746-750.
5. Janssens, R., et al., Patient Preferences in the Medical Product Life Cycle: What do Stakeholders Think? Semi-Structured Qualitative Interviews in Europe and the USA. Patient, 2019. **12**(5): p. 513-526.
6. U. S. Food and Drug Administration, Patient preference information—voluntary submission, review in premarket approval applications, humanitarian device exemption applications, and de novo requests, and inclusion in decision summaries and device labeling. 2016. Google Scholar, 2016.
7. van Overbeeke, E., et al., Factors and situations influencing the value of patient preference studies along the medical product lifecycle: a literature review. Drug Discov Today, 2019. **24**(1): p. 57-68.
8. Brennan, P.F. and I. Strombom, Improving health care by understanding patient preferences: the role of computer technology. J Am Med Inform Assoc, 1998. **5**(3): p. 257-62.
9. Whichello, C., et al., An overview of critical decision-points in the medical product lifecycle: Where to include patient preference information in the decision-making process? Health Policy, 2020. **124**(12): p. 1325-1332.
10. de Bekker-Grob, E.W., et al., Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. Patient, 2015. **8**(5): p. 373-84.
11. Street, D.J., et al., Designing Discrete Choice Experiments for Health Care, in Using Discrete Choice Experiments to Value Health and Health Care. 2008, Springer Netherlands. p. 47-72.
12. Watson, V., F. Becker, and E. de Bekker-Grob, Discrete Choice Experiment Response Rates: A Meta-analysis. Health Econ, 2017. **26**(6): p. 810-817.
13. DeShazo, J.R. and G. Fermo, Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. Journal of Environmental Economics and Management, 2002. **44**(1): p. 123-143.
14. Whichello, C., et al., Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison. BMC Med Inform Decis Mak, 2020. **20**(1): p. 114.
15. Janssens, R., et al., Opportunities and challenges for the inclusion of patient preferences in the medical product life cycle: a systematic review. BMC Med Inform Decis Mak, 2019. **19**(1): p. 189.
16. Berry, S.A., et al., Developing interactions with industry in rare diseases: lessons learned and continuing challenges. Genet Med, 2020. **22**(1): p. 219-226.
17. Watts, S. and P. Stenner, Doing Q methodological research: Theory, method & interpretation. 2012: Sage.
18. Shemmings, D., 'Quantifying' qualitative data: an illustrative example of the use of Q methodology in psychosocial research. Qualitative Research in Psychology, 2006. **3**(2): p. 147-165.

19. Watts, S. and P. Stenner, Doing Q ethodology: theory, method and interpretation. Qualitative Research in Psychology, 2005. **2**(1): p. 67-91.

20. Churruca, K., et al., A scoping review of Q-methodology in healthcare research. BMC Med Res Methodol, 2021. **21**(1): p. 125.

21. van Exel, J., et al., Public views on principles for health care priority setting: findings of a European cross-country study using Q methodology. Soc Sci Med, 2015. **126**: p. 128-37.

22. Jimenez-Moreno, A.C., et al., A study protocol for quantifying patient preferences in neuromuscular disorders: a case study of the IMI PREFER Project. Wellcome Open Res, 2020. **5**: p. 253.

23. Deenen, J.C., et al., The Epidemiology of Neuromuscular Disorders: A Comprehensive Overview of the Literature. J Neuromuscul Dis, 2015. **2**(1): p. 73-85.

24. Katirji, B., Clinical Assessment in Neuromuscular Disorders, in Neuromuscular Disorders in Clinical Practice. 2014, Springer New York. p. 3-20.

25. D'Angelo, M.G. and N. Bresolin, Cognitive impairment in neuromuscular disorders. Muscle Nerve, 2006. **34**(1): p. 16-33.

26. Bresolin, N., et al., Cognitive impairment in Duchenne muscular dystrophy. Neuromuscul Disord, 1994. **4**(4): p. 359-69.

27. Astrea, G., et al., Reading impairment in Duchenne muscular dystrophy: A pilot study to investigate similarities and differences with developmental dyslexia. Res Dev Disabil, 2015. **45-46**: p. 168-77.

28. Cotton, S., N.J. Voudouris, and K.M. Greenwood, Intelligence and Duchenne muscular dystrophy: Full-Scale, Verbal, and Performance intelligence quotients. Developmental Medicine & Child Neurology, 2007. **43**(7): p. 497-501.

29. Landfeldt, E., et al., Disease burden of myotonic dystrophy type 1. J Neurol, 2019. **266**(4): p. 998-1006.

30. Gorman, G.S., et al., Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. Ann Neurol, 2015. **77**(5): p. 753-9.

31. Johnson, N.E., et al., Population-Based Prevalence of Myotonic Dystrophy Type 1 Using Genetic Analysis of Statewide Blood Screening Program. Neurology, 2021. **96**(7): p. e1045-e1053.

32. Stephenson, W., Correlating persons instead of tests. Character & Personality; A Quarterly for Psychodiagnostic & Allied Studies, 1935.

33. Stephenson, W., Technique of Factor Analysis. Nature, 1935. **136**(3434): p. 297-297.

34. Jimenez-Moreno, A.C., et al., Patient Preferences in Rare Diseases: A Qualitative Study in Neuromuscular Disorders to Inform a Quantitative Preference Study. Patient, 2021. **14**(5): p. 601-612.

35. Brown, J.D., Choosing the right type of rotation in PCA and EFA. JALT testing & evaluation SIG newsletter, 2009. **13**(3): p. 20-25.

36. Guttman, L., Some necessary conditions for common-factor analysis. Psychometrika, 1954. **19**(2): p. 149-161.

37. Kaiser, H.F., The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement, 2016. **20**(1): p. 141-151.

38. Cohen, J., Statistical power analysis for the behavioral sciences. 2013: Academic press.

39. Akhtar-Danesh, N., Qfactor: A Command for Q-methodology Analysis. The Stata Journal: Promoting communications on statistics and Stata, 2018. **18**(2): p. 432-446.

40. Wenninger, S., F. Montagnese, and B. Schoser, Core Clinical Phenotypes in Myotonic Dystrophies. Front Neurol, 2018. **9**: p. 303.

41. Walklet, E., et al., Do Psychosocial Interventions Improve Quality of Life and Wellbeing in Adults with Neuromuscular Disorders? A Systematic Review and Narrative Synthesis. J Neuromuscul Dis, 2016. **3**(3): p. 347-362.

42. Miro, J., et al., Chronic pain in neuromuscular disease: pain site and intensity differentially impacts function. Phys Med Rehabil Clin N Am, 2012. **23**(4): p. 895-902.

43. Abresch, R.T., et al., Assessment of pain and health-related quality of life in slowly progressive neuromuscular disease. Am J Hosp Palliat Care, 2002. **19**(1): p. 39-48.

44. Jensen, M.P., et al., Chronic pain in persons with myotonic dystrophy and facioscapulohumeral dystrophy. Arch Phys Med Rehabil, 2008. **89**(2): p. 320-8.

45. Willig, T.N., et al., Swallowing problems in neuromuscular disorders. Arch Phys Med Rehabil, 1994. **75**(11): p. 1175-81.

46. LaDonna, K.A., W.J. Koopman, and S.L. Venance, Myotonic dystrophy (DM1) and dysphagia: the need for dysphagia management guidelines and an assessment tool. Can J Neurosci Nurs, 2011. **33**(1): p. 42-6.

47. Read, J.L., et al., Prevalence and severity of voice and swallowing difficulties in mitochondrial disease. Int J Lang Commun Disord, 2012. **47**(1): p. 106-11.

48. Truijens, D. and J. van Exel, Views on deceased organ donation in the Netherlands: A q-methodology study. PLoS One, 2019. **14**(5): p. e0216479.

49. Duncan Millar, J., H. Mason, and L. Kidd, What is Q methodology? Evid Based Nurs, 2022. **25**(3): p. 77-78.

50. Corr, S., R. Capdevila, and C. Phillips, Using Q methodology to evaluate a day service for younger adult stroke survivors. Operant Subjectivity: The International Journal of Q Methodology, 2003. **27**(1): p. 1-23.

51. Westbrook, J.L., et al., Validity of dementia care mapping on a neuro-rehabilitation ward: Q-methodology with staff and patients. Disabil Rehabil, 2013. **35**(19): p. 1652-9.

Chapter 5

Does it matter how you ask? Assessing the Impact of Failure or Effectiveness Framing on Preferences for Antibiotic Treatments in a Discrete Choice Experiment

Authors: Ian P. Smith*, Mirko Ancillotti*, Esther W. de Bekker-Grob, Jorien Veldwijk

*Shared first author

Status: Published

Smith, I.P., Ancillotti, M., de Bekker-Grob, E.W. and Veldwijk, J., 2022. Does It Matter How You Ask? Assessing the Impact of Failure or Effectiveness Framing on Preferences for Antibiotic Treatments in a Discrete Choice Experiment. Patient Preference and Adherence, pp.2921-2936.

ABSTRACT

Purpose: Studies assessing framing effects in Discrete Choice Experiments (DCE) primarily focused on attributes related to mortality/survival information. Little is known about framing effects for other attributes in health related DCEs. This study aimed to investigate how framing treatment outcome as effective, failure, or a combined frame impacts respondent choices and DCE outcomes.

Patients and methods: Three Bayesian D-efficient designed DCE surveys measuring preferences for antibiotic treatments were randomly distributed to a representative sample of the Swedish population aged 18-65 years (n=1119). Antibiotic treatments were described using five attributes. Four attributes were static: Contribution to Antibiotic Resistance, Treatment Duration, Likelihood of Side-Effects, and Costs. A fifth treatment attribute was framed in three ways: Effectiveness, Failure Rate, or both. Mixed logit models were used to analyze attribute level estimates, importance value, and choice predictions.

Results: Significant differences between the frames were found for the parameter estimates of the attributes of Treatment Duration and Likelihood of Side-Effects, but not Treatment Outcome which was the alternatively framed attribute. Contribution to Antibiotic Resistance and Costs were the most important attributes for all participants regardless of framing. Choice predictions for the "best option" antibiotic only slightly differed between the groups based on the frame seen (95.2-92.4%).

Conclusion: Our study showed that attribute framing can impact preferences regardless of the attribute's importance value in alternative valuation. However, the practical implication of this effect may be limited. A theoretical discussion is needed to identify how researchers should accommodate and report any potential framing effect in their studies.

INTRODUCTION

Discrete choice experiments (DCEs) are a commonly used method to elicit stated preferences for health treatment outcomes.[1-3] In a DCE, participants are asked to choose between different choices or alternatives, which are described using sets of pre-specified attributes and levels. The foundational theory of DCEs, the Random Utility Theory (RUT), assumes that the utility of an alternative can be derived through the compound valuation of the different attributes and levels describing the alternative.[4-6] This total utility is used by the participant to compare alternatives in a choice context and the resulting choice should represent the greatest total utility according to RUT.[7]

When developing a DCE, researchers are tasked with creating a limited set of attributes to describe the alternatives. This set of attributes should be based on a thorough qualitative process including a literature review, interviews with experts, and discussions with the (potential) consumers of the healthcare product.[8] The goal of this process is to create a list of attributes which is thorough enough to contain those attributes which are relevant to the consumers of the healthcare product and stakeholders who will use this information without being so large as to make the study design over complex or burdensome.[9-11] A key part of this process is having attributes described in such a way that they can be clearly understood by the participants. In order to ensure that the attributes are understood, qualitative assessment of the attributes by the researchers is recommended by good practice guidelines, but often limited to checks of comprehensibility and plausibility.[9, 12] Current guidelines for developing an attribute list in patient preference studies do not address the impact that framing can have on interpreting attributes.[9, 13-15]

The impact that framing can have on decision-making is well documented and has been studied for years in psychological fields.[16-18] This framing bias has been found to influence a participant's understanding of the health information commonly used in DCEs, such as probability of negative outcomes or treatment risks.[19-22] Specific examples include patients preferring lung cancer treatments framed in terms of survival rather than mortality [23], an effect so great that the framing can move an attribute from being the least important to the most important when eliciting preferences.[24] In one study participants were more likely to choose a healthy lifestyle when the outcomes were framed as an increase in life expectancy vs poor health prevention.[25] In another study, participants were more willing to pay for colorectal screening when the tests were framed in regards to percentage of cancers found versus missed.[26] Most of these studies are similar in that they assessed the impact of framing for attributes that generally play a large role in decision making (such as survival or mortality). In the field of healthcare preference assessment, only the study by Howard and Salkeld looked at the impact attribute framing had on an attribute not related to survival or mortality (i.e. specificity or sensitivity of colorectal cancer screening).[26] None of the studies presented the different frames simultaneously as a means to possibly control for the framing effect which

has previously been suggested as a method to control for framing effects.[27] There is thus a lack of research specific to the impact of framing in the field of patient preference assessment and the best practices to overcome it.[28]

This article reports on a methodological study designed specifically to address this gap in knowledge by assessing the impact of framing on preferences for a hypothetical personal medical treatment. Specifically, this study aims to assess whether presenting an attribute not related to survival or mortality in a positive, negative, or combined frame impacts preference outcomes in terms of stated choices, preference estimates, importance value scores, and expected choice predictions.

METHODS

Case study and participants

A DCE was developed to assess preferences for different antibiotic treatments when considering antibiotic resistance as an attribute of this treatment.[29] Antibiotics are a commonly prescribed medical treatment for bacterial infection and, like any medical treatment, they are often described in regards to treatment outcome making them an ideal subject for this preference task.[30-32] The participant group was recruited via Dynata, a commercial survey panel provider. The sample was drawn from the general Swedish population and was nationally representative in terms of age, gender, education, and geographic region. The inclusion criteria for the study were that the participant was 18-65 years of age, proficient in Swedish language, and self-reported being medically able to take antibiotics. Respondents were excluded if they could not take antibiotics (eg, due to allergies). Participants were recruited until the desired sample size (N=350 in each arm) was reached as this should provide sufficient power for the analysis.[33] Informed consent was obtained from all participants before starting and prior to completing the survey.

This study was planned in adherence to Swedish research regulations and was evaluated and approved by the Uppsala Regional Ethical Review Board (Dnr 2018/293) and complies with the Declaration of Helsinki.

Attributes, levels and experimental design

The antibiotic treatments were described using five different attributes: *Contribution to Antibiotic Resistance*, *Duration of Treatment*, *Likelihood of Side-Effects*, *Treatment Outcome*, and *Cost*. The attributes were developed using a qualitative process involving literature reviews and four focus groups including ranking exercises with members of the general population (13 women/10 men, mean age = 38 years, age range 20–81 years). The attributes generated were then reduced and refined in discussion with research colleagues in accordance with best practice guidelines.[9, 14, 29] The levels for the attributes *Duration of Treatment*, *Likelihood*

of Side-Effects, *Treatment Outcome*, and Cost were derived based on characteristics of the most commonly prescribed antibiotics used in clinical care in Sweden. Costs were included to reflect the co-pay insurance system where patients pay up to a fixed amount for medications each year which is the standard in Sweden. The levels for *Contribution to Antibiotic Resistance* were derived based on the report from the Swedish public health authority regarding current levels of antibiotic resistance.[34, 35] The final list of attributes and levels was reviewed in eight stakeholder interviews (N=4 patients, N=1 expert on antibiotic resistance, N=1 nurse, N=2 general practitioners). All attributes were described identically for all participants except for *Treatment Outcome*.

In order to address the research question, three DCEs with different frames were designed for the attribute of *Treatment Outcome*. *Treatment Outcome* was framed as Failure Rate (Arm 1), Effectiveness (Arm 2), or both Effectiveness and Failure Rate combined (Arm 3). The levels for this attribute were logically identical across the three study arms, while being framed in opposite terms. For example, if in Arm 1 the Failure Rate is described as "5%", then in Arm 2 the Effectiveness is "95%", and in Arm 3 "95%/5%". Thus, all treatment options were fundamentally the same even if framed in different ways. The attributes included can be found in Table 1 along with the full description presented to the participant and the levels for each attributed.

For each Arm frame, the DCE design was optimized using a Bayesian D-efficient design created with Ngene 1.0 (ChoiceMetrics, 2011). The initial prior preference information was generated using best guess estimates based on literature review and expert opinions (N=1 pharmacist, N=1 nurse, N=2 general practitioners). The choice tasks consisted of a forced-choice in which a participant had to choose between two unlabeled alternatives ('Antibiotic A' or 'Antibiotic B'). The decision to use a forced choice instead of an unforced choice with an opt-out alternative was based on the methodological needs of this study and desire to increase the efficiency of the design.[36] All participants were asked to imagine that they were planning to take antibiotics to treat a bacterial infection. The two alternatives were presented as choices the doctor gave them to treat the infection. An example choice task can be found in Figure 1.

A pilot test of the complete survey was conducted in February 2019. The pilot test consisted of 129 respondents evenly split between the different arms and recruited using the same methods and research population as the final survey. Coefficients to be used as Bayesian priors for the experimental design of the final DCE were derived from the output of multinomial logit (MNL) models fitted to the pilot data. The final Bayesian D-efficient design consisted of 48 unique choice tasks divided over 3 blocks of 16 choice tasks for each arm (respondents were randomly assigned to these blocks).

Table 1. Attributes and Levels for each Attribute used in the choice tasks along with explanatory text as presented to the participants

Attribute (Description)	**Levels**			
Contribution to antibiotic resistance (static, i.e. presented to all respondents)				
Antibiotic resistance results in higher costs for care and treatment. In case of infections caused by resistant bacteria, effective treatment is delayed. Thus, the healing time is prolonged and the risk of complications increases.				
Primary Text	Low	Medium	High	
Hoverbox Text	15,000 cases each year: in the next 10 years the number of cases in Sweden remains the same	30,000 cases each year: in the next 10 years the number of cases in Sweden doubles	70,000 cases each year: in the next 10 years the number of cases in Sweden more than quadruples	
Treatment duration (static)				
The treatment time as prescribed by the doctor.				
	3 days	7 days	14 days	
Likelihood of Side-Effects (static)				
Antibiotics have, as all medicines, Side-Effects. As they not only kill the harmful but also the beneficial bacteria you have in your body, they can give mild to moderate Side-Effects such as nausea, upset stomach, headache and fatigue.				
Primary Text	1%	5%	10%	20%
Hoverbox Text	1 out of 100 people taking the antibiotic experiences Side-Effects, 99 do not experience Side-Effects	5 out of 100 people taking the antibiotic experiences Side-Effects, 95 do not experience Side-Effects	10 out of 100 people taking the antibiotic experiences Side-Effects, 90 do not experience Side-Effects	20 out of 100 people taking the antibiotic experiences Side-Effects, 80 do not experience Side-Effects
Treatment Outcome: Failure Rate (presented to 33.7% of respondents)				
Failure of an antibiotic treatment is the extent to which the antibiotic fails in its intended effect: to treat the infection. Not all treatments are equally effective, if an antibiotic treatment fails, you must be treated with another antibiotic.				
Primary Text	5%	10%	15%	20%
Hoverbox Text	5 out of 100 people need an additional antibiotic cure	10 out of 100 people need an additional antibiotic cure	15 out of 100 people need an additional antibiotic cure	20 out of 100 people need an additional antibiotic cure

(cont'd)

Treatment Outcome: Effectiveness (presented to 32.9% of respondents)				
Effectiveness of an antibiotic treatment is the extent to which the antibiotic achieves its intended effect: to treat the infection. Not all treatments are equally effective, if an antibiotic treatment is not effective, you must be treated with another antibiotic.				
Primary Text	95%	90%	85%	80%
Hoverbox Text	95 out of 100 people heal from infection	90 out of 100 people heal from infection	85 out of 100 people heal from infection	80 out of 100 people heal from infection
Treatment Outcome: Effectiveness / Failure Rate (presented to 33.4% of respondents)				
An antibiotic treatment is effective when it treats the infection and the patient recovers. Sometimes the treatment fails for several different reasons. If the treatment is not effective, you will need to be treated with another course of another antibiotic.				
Primary Text	95%/5%	90%/10%	85%/15%	80%/20%
Hoverbox Text	95 out of 100 people heal from infection 5 out of 100 people need an additional course of antibiotics	90 out of 100 people heal from infection 10 out of 100 people need an additional course of antibiotics	85 out of 100 people recover from the infection 15 out of 100 people need an additional course of antibiotics	80 out of 100 people recover from the infection 20 out of 100 people need an additional course of antibiotics
Cost for you*				
These antibiotics are not subsidized via the high-cost protection / drug benefits, so you have to pay the full cost yourself.				
	€10	€25	€40	€100

Note: * Static: presented to all respondents; ₮ Currency was presented in Swedish Krona to participants but converted to Euros for reporting using a rounded April 2019 exchange rate 10 krona/ 1 euro

Questionnaire

The full questionnaire consisted of three different sections. The first section assessed demographic information such as age, gender, highest attained educational level, and occupation. In addition to these, two validated sets of questions were used to determine respondent's health literacy (the Communicative and Critical Health Literacy Scale – Swedish Version; S-CCHL) [37] and numeracy (the 3-Item Version of the Subjective Numeracy Scale; SNS-3).[38] Health literacy and numeracy reflect the participants ability to utilize health information presented in text form or numeric form (respectively). The S-CCHL assesses health literacy using five items on a five-point Likert scale ('never (1)' to 'always (5)'. The SNS-3 consists of three items on a six-point Likert scale from 'not good at all/never (1)' to 'extremely good/very often (6)'. Overall scores for each scale were generated by averaging their responses. Participants were then categorized as follows: scores 1/2 were classed as 'inadequate'; those who had at least one score of 3 in the S-CCHL and 3/4 in the SNS-3 were classed as 'problematic'; and those who consistently scored 4/5 in the S-CCHL and 5/6 in the SNS-3 were classed

as 'sufficient'. These scores were used to ensure the comparability of the samples in regard to their ability comprehend written and numerical health information. Finally, six questions were used to assess experience with antibiotics, knowledge about antibiotic resistance, and self-reported health status.

The second section of the survey consisted of the DCE. Participants were instructed to imagine that they had a non-life-threatening bacterial infection and their doctor wants to prescribe antibiotics to treat the infection and avoid complications. Participants were randomized to receive a survey with *Treatment Outcome* framed as either treatment effectiveness, treatment failure, or a combination of both. Prior to answering the choice tasks, respondents received descriptions of bacterial infections, antibiotic treatments and antibiotic resistance. Additional information explaining the attributes and levels was available in a pop-up window which the participant could view during the task by hovering over the attribute. The attributes of *Likelihood of Side-Effects* and *Treatment Outcome* were presented as a percentage in the choice task and as a text and an icon array of these percentages in a pop-up box for further explanation to ensure participant comprehension.[19] See Figure 1 for an example choice task with pop-up window.

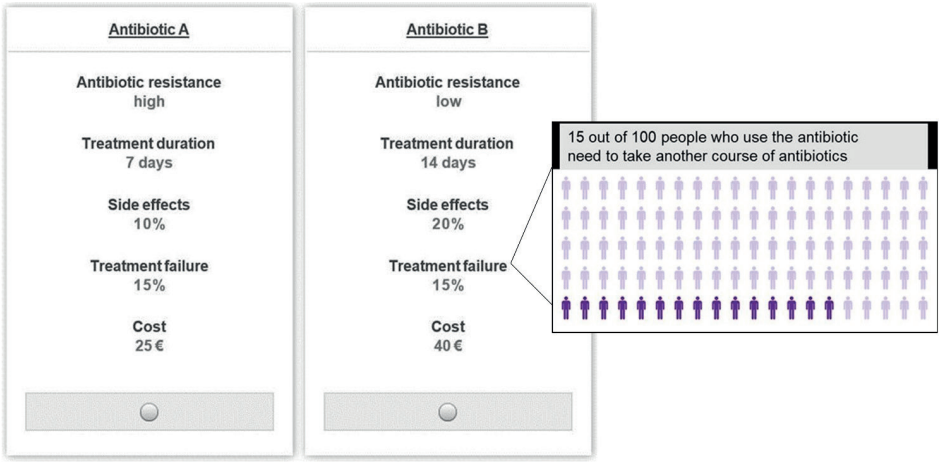


Figure 1 Example choice task using a treatment failure framing with pop-up window

The final section consisted of short questions that assessed the participant's evaluation of the questionnaire in regard to length and difficulty on a 5-point Likert scale along with an optional comments field. The questionnaire was developed using Light House Studio 9.6.1 software. The entire survey was pre-tested test to assess comprehensibility by eight stakeholders. This pre-test resulted in two changes: rewording a knowledge related question, and the addition of an exclusion criterion regarding ability to use antibiotics.

Statistical Analysis

All analyses were conducted using STATA 14 clogit, and the mixlogit packages. A significance level of 0.05 was used for all analyses. The reference case used for all analyses was an antibiotic treatment with low contribution to antibiotic resistance, a 3-day treatment duration, 1% likelihood of side-effects, high effectiveness/low failure rate/ (95%/5%), and cost of €10.

A straight-line test was done to see if participants only chose either the left or right alternative. Potential scale parameter differences between data of the three arms were assessed using a Swait and Louviere test.[39] Three tests were done comparing the three model pairs (Failure-Effectiveness, Failure-Combined, Effectiveness-Combined) following the same procedure: the log likelihood of the MNL model was fitted separately in all three frames and those were tested against the log likelihood of the MNL model for the different pooled data sets, which accounted for potential scale parameter differences. The hypothesis of equal attribute level estimates was rejected in all comparisons by means of the chi-squared test.

Parameter coefficient estimates and importance value scores for the different arms were then derived using a two-step process. First, a logit model was used to identify significant main effects, interaction effects, linearity of parameter estimates across attribute levels, and left-right bias.[40] No attributes were found to be linear across all three arms thus each attribute level was added to the model as a single parameter using dummy coding to allow for easy interpretation of results and willingness-to-pay estimates were not possible to calculate.[41] All possible attribute level/framing-arm interaction terms were checked in an initial logit model (e.g. Effectiveness* Duration: 7 days; Failure* Duration: 7 days; Combination* Duration: 7 days). Interaction terms that did not significantly add to the model were removed using backwards elimination. A mixed logit model (MIXL) was used for this analysis as it allows for the inclusion of random effects parameters accounting for naturally occurring heterogeneity of preferences in samples and results in more accurate model estimations.[42] A MIXL model was built for each individual arm as well as the entire population in order to show the preference differences and test for interaction effects. All attributes were found to have significant heterogeneity at some level and were thus all included as random parameters. Each model was built using 14,000 Halton draws to ensure robust results.[43] The final MIXL model can be found in Equation 1.

Equation 1.

$$U_{iac} = \beta_{1i}Contribution\ AR_{medium} + \beta_{2i}Contribution\ AR_{high} + \beta_{3i}T_xDuration_{\ 7\ days} + \beta_{4i}T_xDuration_{\ 14\ days} + \beta_{5i}Side\ Effects_{\ 5\%} + \beta_{6i}Side\ Effects_{\ 10\%} + \beta_{7i}Side\ Effects_{\ 20\%} + \beta_{8i}T_xOutcome_{\ 90\%/10\%} + \beta_{9i}T_xOutcome_{\ 85\%/15\%} + \beta_{10i}T_xOutcome_{\ 80\%/20\%} + \beta_{11i}Costs_{\ €25} + \beta_{12i}Costs_{\ €40} + \beta_{13i}Costs_{\ €100} + \beta_{14a_i}T_xDuration_{\ 7\ days} + \beta_{15a_i}T_xDuration_{\ 14\ days} + \beta_{16a_i}Side\ Effects_{\ 5\%} + \beta_{17a_i}Side\ Effects_{\ 10\%} + \beta_{18a_i}Side\ Effects_{\ 20\%} + \varepsilon_{ijc}$$

$$U_{ibc} = \beta_{1i}Contribution\ AR_{medium} + \beta_{2i}Contribution\ AR_{high} + \beta_{3i}T_xDuration_{\ 7\ days} + \beta_{4i}T_xDuration_{\ 14\ days} + \beta_{5i}Side\ Effects_{\ 5\%} + \beta_{6i}Side\ Effects_{\ 10\%} + \beta_{7i}Side\ Effects_{\ 20\%} + \beta_{8i}T_xOutcome_{\ 90\%/10\%} + \beta_{9i}T_xOutcome_{\ 85\%/15\%} + \beta_{10i}T_xOutcome_{\ 80\%/20\%} + \beta_{11i}Costs_{\ €25} + \beta_{12i}Costs_{\ €40} + \beta_{13i}Costs_{\ €100} + \beta_{14a_i}T_xDuration_{\ 7\ days} + \beta_{15a_i}T_xDuration_{\ 14\ days} + \beta_{16a_i}Side\ Effects_{\ 5\%} + \beta_{17a_i}Side\ Effects_{\ 10\%} + \beta_{18a_i}Side\ Effects_{\ 20\%} + \varepsilon_{ijc}$$

In this model, the value (U) of alternative a or b in a specific decision context (c) for an individual (i) is derived as the sum of the attribute-level estimates indicating the importance value of each attribute level ($\beta_1 - \beta_{13}$), plus the specific parameter interaction term coefficients for the arm (a) of that individual ($\beta_{14} - \beta_{18}$). The stochastic factors for this alternative are included in the utility function as a random error term ε. Importance value scores (IVS) and choice predictions were derived from the results of the MIXL models. The IVS were generated by standardizing the score of the attribute with the greatest absolute utility difference between the least preferred level and most preferred level to 1.[44] The attribute with the greatest absolute utility for the Failure and the Effectiveness arms was *Contribution to Antibiotic Resistance* and for the combination arm it was *Costs*.

In order to compare the differences as they may be used in an applied setting, hypothetical choice predictions were calculated as proportion of exponentiated utility attributed to an alternative from the total exponentiated utility present in a choice scenario (see equation 2). In this equation, the uptake probability is calculated as the mean uptake of alternative (V) for an individual when asked to choose between this alternative or another (W). The choice predictions presented show the likely uptake if a participant was presented with either the most desirable anti-biotic for their specific frame, the least desirable anti-biotic for their specific frame, or an Amoxicillin proxy. Amoxicillin was chosen as it is one of the most commonly prescribed antibiotics for out-patient usage.[45] For the purposes of this calculation, Amoxicillin was described based on descriptions found in published literature. This description is having a high contribution to antibiotic resistance[46], a treatment duration of 7 days, 10% of users experiencing a side effect[47], an 85%/15% effectiveness/failure rate[48], with low costs.[49] In addition to this, individual probabilities were generated per person based on the individual parameter estimates to see what the likely individual choice predictions would be.

Equation 2.

$$\sum_{i=1}^{n} \frac{e^{V_i}}{e^{V_i} + e^{W_i}}$$

RESULTS

Study Population

In total, 1124 completed responses were received. A straight-line test resulted in 5 responses being excluded due to selecting all right or all left alternatives, thus the final analysis included 1119 responses. A summary of respondent demographic information can be found in Table 2. Participants generally found the survey acceptable in regards to length (M=2.72, SD=5.94) and not difficult (M= 3.42, SD=.859) with no significant differences found between the groups (F(2)= 0.145, p=0.865) and F(2)=0.71, p = 0.932 respectively. The percentage of people who showed dominant decision-making, in which their choices were always associated with the

best level of one specific attribute, did not significantly differ across arm and was 19%, 22% and 23% for the effectiveness, failure rate and the combined arm respectively. In all arms, these lexicographical preferences were mostly registered for the resistance attribute and the cost attribute.

Table 2 Respondent Demographic Information

		Survey 1: Failure N=377	**Survey 2: Effective N=368**	**Survey 3: Combination N=374**
Age M (SD, range)		43.23 (13.529, 18-65)	43.10 (13.940, 18-65)	42.61 (13.581, 18-65)
Female N (%)		208 (55.2)	204 (55.4)	196 (52.4)
Education Level N (%) Ŧ*	EQF 1-2	23 (6.2)	37 (10.2)	30 (8.1)
	EQF 3	46 (12.4)	71 (19.6)	52 (14.1)
	EQF 4-5	108 (29.1)	109 (30.)	109 (29.5)
	EQF 6	96 (25.9)	72 (19.8)	103 (27.8)
	EQF 7-8	98 (26.4)	74 (20.4)	76 (20.5)
Occupation N (%)				
	Employed (permanent, temporary, self-employed)	248 (65.8)	234 (63.6)	220 (58.8)
	Students	36 (9.6)	37 (10.1)	44 (11.8)
	Retired	33 (8.6)	41 (11.1)	40 (10.7)
	Unemployed	41 (10.9)	29 (7.9)	49 (13.1)
	On disability living allowance, sick leave and other	19 (5.0)	27 (7.3)	21 (5.6)
Health Literacy N (%)				
	Low	41 (10.9)	47 (12.7)	47 (12.6)
	Medium	161 (42.7)	146 (40.5)	140 (37.4)
	High	175 (46.4)	172 (46.7)	187 (50.0)
Numeracy N (%)				
	Low	107 (28.4)	103 (28.0)	104 (27.8)
	Medium	182 (48.3)	185 (50.3)	188 (50.3)
	High	88 (23.3)	80 (21.7)	82 (21.9)
Antibiotic experience N (%)				
	Yes	332 (88.1)	336 (91.3)	322 (86.1)
	Never	19 (5.0)	20 (5.4)	22 (5.9)
	Don't know	26 (6.9)	12 (3.3)	30 (8.0)

Note: Ŧ Education was measured as the European Qualifications Framework (EQF) level; * Significant differences between arms at $p<0.05$

Preference Estimates

The preference estimates resulting from the MIXLs for each individual arm can be found in Table 3. All attributes were found to have significant parameter estimates for at least one level of each arm. Significant heterogeneity was found for at least one level of each parameter in the MIXL model in each arm indicating that the samples participants were highly varied in how they valued the different attributes when assessing antibiotics. The exception to this was the attribute of *Treatment Outcome* in the Effectiveness arm.

In the pooled sample, no significant interaction effect was found for the attribute of *Treatment Outcome* despite this being the attribute which was framed differently between the arms. Significant interaction effects were found between the way that *Treatment Effectiveness* was framed and the attributes of *Likelihood of Side-Effects* and *Treatment Duration*. The *Likelihood of Side-Effects* was significantly more important to participants who saw the effectiveness framing compared to respondents who saw the Failure or combined framing. For *Treatment Duration*, those who saw a Failure Rate frame used this attribute to a greater extent than Effectiveness or Combined Frame. Further, they viewed a treatment lasting 7 days as somewhat better than a duration of 3 or 14 days. The preference estimates resulting from the MIXLs for all arms together with the significant framing interaction terms can be found in Table 4.

Table 3. MIXL Model results for the three different arms

		Survey 1: Failure Framing					Survey 2: Effectiveness Framing					Survey 3: Failure and Effectiveness Frames				
				95% CI					95% CI					95% CI		
Attribute: Level		**β-coeff**	**SE**	**Lower**	**Upper**	*P*	**β-coeff**	**SE**	**Lower**	**Upper**	*P*	**β-coeff**	**SE**	**Lower**	**Upper**	*P*
Contribution to Resistance: Low	(ref)															
Contribution to Resistance: Med	Mean	-0.9913	0.0861	-1.1601	-0.8226	<0.001	-1.0338	0.1037	-1.2370	-0.8306	<0.001	-0.9546	0.0950	-1.1408	-0.7684	<0.001
	SD	-0.7150	0.1074	-0.9256	-0.5045	<0.001	1.2542	0.1180	1.0228	1.4855	<0.001	0.9426	0.1024	0.7419	1.1432	<0.001
Contribution to Resistance: High	Mean	-2.6755	0.2000	-3.0674	-2.2836	<0.001	-2.6670	0.2062	-3.0711	-2.2630	<0.001	-2.6057	0.2021	-3.0019	-2.2095	<0.001
	SD	2.4438	0.1674	2.1157	2.7719	<0.001	2.6875	0.1942	2.3069	3.0681	<0.001	2.6684	0.1882	2.2994	3.0373	<0.001
T_X Duration: 3 days	(ref)															
T_X Duration: 7 days	Mean	0.1455	0.0659	0.0163	0.2747	0.027	-0.0089	0.0705	-0.1471	0.1292	0.899	-0.0236	0.0685	-0.1579	0.1106	0.730
	SD	-0.0431	0.2324	-0.4987	0.4125	0.853	-0.3488	0.1502	-0.6432	-0.0545	0.020	0.0501	0.1492	-0.2422	0.3425	0.737
T_X Duration: 14 days	Mean	-0.4154	0.0744	-0.5611	-0.2696	<0.001	-0.3295	0.0748	-0.4762	-0.1829	<0.001	-0.4547	0.0749	-0.6014	-0.3080	<0.001
	SD	-0.5871	0.0892	-0.7620	-0.4122	<0.001	-0.3460	0.1463	-0.6329	-0.0592	0.018	-0.5708	0.0984	-0.7637	-0.3779	<0.001
Likelihood of Side-Effects: 1%																
Likelihood of Side-Effects: 5%	Mean	-0.3045	0.0880	-0.4770	-0.1319	0.001	-0.6217	0.0855	-0.7892	-0.4542	<0.001	-0.2715	0.0928	-0.4535	-0.0895	0.003
	SD	-0.2637	0.2468	-0.7475	0.2201	0.285	-0.0002	0.2128	-0.4173	0.4170	0.999	0.3210	0.2501	-0.1693	0.8112	0.199
Likelihood of Side-Effects: 10%	Mean	-0.4389	0.0845	-0.6045	-0.2733	<0.001	-1.0859	0.0945	-1.2711	-0.9006	<0.001	-0.7603	0.0997	-0.9557	-0.5648	<0.001
	SD	0.1698	0.2123	-0.2463	0.5859	0.424	-0.1348	0.5007	-1.1161	0.8465	0.788	-0.2128	0.2407	-0.6846	0.2590	0.377
Likelihood of Side-Effects: 20%	Mean	-1.0981	0.1192	-1.3318	-0.8644	<0.001	-1.7286	0.1357	-1.9945	-1.4627	<0.001	-1.3655	0.1384	-1.6368	-1.0943	<0.001
	SD	1.0129	0.1109	0.7954	1.2303	<0.001	1.2727	0.1312	1.0156	1.5298	<0.001	1.2574	0.1281	1.0063	1.5085	<0.001

(cont'd)

		Survey 1: Failure Framing					Survey 2: Effectiveness Framing					Survey 3: Failure and Effectiveness Frames				
				95% CI					95% CI					95% CI		
Attribute: Level		**β-coeff**	**SE**	**Lower**	**Upper**	*P*	**β-coeff**	**SE**	**Lower**	**Upper**	*P*	**β-coeff**	**SE**	**Lower**	**Upper**	*P*
T_X Effectiveness/Failure: 95%/5%	(ref)															
T_X Effectiveness/Failure: 90%/10%	Mean	-0.3778	0.1356	-0.6435	-0.1121	0.005	-0.1923	0.1803	-0.5457	0.1612	0.286	0.1481	0.2841	-0.4087	0.7050	0.602
	SD	0.6172	0.1930	0.2389	0.9954	0.001	0.2578	0.3639	-0.4554	0.9709	0.479	0.5386	0.1862	0.1736	0.9036	0.004
T_X Effectiveness/Failure: 85%/15%	Mean	-0.8237	0.1479	-1.1136	-0.5338	<0.001	-0.3279	0.2000	-0.7198	0.0641	0.101	-0.5427	0.2805	-1.0924	0.0071	0.053
	SD	0.3907	0.2794	-0.1569	0.9382	0.162	-0.3083	0.3792	-1.0515	0.4349	0.416	-0.2530	0.3208	-0.8818	0.3759	0.430
T_X Effectiveness/Failure: 80%/20%	Mean	-1.0964	0.1572	-1.4045	-0.7882	<0.001	-0.8442	0.1923	-1.2212	-0.4672	<0.001	-0.7732	0.0956	-0.9605	-0.5859	<0.001
	SD	1.0496	0.1407	0.7739	1.3253	<0.001	-0.4985	0.3851	-1.2532	0.2562	0.195	0.9375	0.1172	0.7078	1.1672	<0.001
Cost: €10	(ref)															
Cost: €25	Mean	-0.2694	0.0861	-0.4381	-0.1008	0.002	-0.2374	0.0836	-0.4012	-0.0735	0.005	-0.2729	0.0792	-0.4282	-0.1176	0.001
	SD	-0.0693	0.2002	-0.4617	0.3230	0.729	0.0360	0.2267	-0.4084	0.4804	0.874	-0.0515	0.1750	-0.3944	0.2915	0.769
Cost: €40	Mean	-0.9274	0.1011	-1.1255	-0.7293	<0.001	-0.6950	0.0897	-0.8708	-0.5192	<0.001	-0.8999	0.1050	-1.1057	-0.6940	<0.001
	SD	-0.6175	0.1274	-0.8673	-0.3677	<0.001	-0.3837	0.2025	-0.7806	0.0132	0.058	0.8021	0.1104	0.5858	1.0184	<0.001
Cost: €100	Mean	-2.5514	0.2230	-2.9885	-2.1143	<0.001	-2.6364	0.2038	-3.0358	-2.2371	<0.001	-2.8258	0.2181	-3.2533	-2.3982	<0.001
	SD	2.7336	0.1913	2.3586	3.1085	<0.001	2.6289	0.2064	2.2243	3.0335	<0.001	2.8260	0.2025	2.4291	3.2230	<0.001
Log-likelihood		-3091.3249					-2783.2115					-3057.4523				

Note: SD Standard deviation; SE Standard Error; T_X Treatment Parameters with significant Random effects are shown with SD

Table 4. Attribute-level estimates for the MIXL Model of the pooled results from all arms

		random effects model				
				95% CI		
Attribute: Level		**β-coeff**	**SE**	**Lower**	**Upper**	*P*
Contribution to Resistance: Low	(ref)					
Contribution to Resistance: Med	Mean	-0.9805	0.0526	-1.0837	-0.8774	<0.001
	SD	0.9669	0.0612	0.8471	1.0868	<0.001
Contribution to Resistance: High	Mean	-2.5970	0.1112	-2.8151	-2.3790	<0.001
	SD	2.5959	0.1059	2.3884	2.8034	<0.001
T_X Duration: 3 days	(ref)					
T_X Duration: 7 days*	Mean	0.1623	0.0649	0.0352	0.2894	0.012
	SD	-0.0424	0.1590	-0.3541	0.2693	0.790
T_X Duration: 14 days*	Mean	-0.3765	0.0691	-0.5119	-0.2411	<0.001
	SD	-0.5161	0.0585	-0.6309	-0.4014	<0.001
Likelihood of Side-Effects: 1%*	(ref)					
Likelihood of Side-Effects: 5%*	Mean	-0.3455	0.0843	-0.5106	-0.1803	<0.001
	SD	-0.1442	0.1605	-0.4588	0.1704	0.369
Likelihood of Side-Effects: 10%*	Mean	-0.4081	0.0773	-0.5596	-0.2566	<0.001
	SD	-0.1898	0.1652	-0.5135	0.1339	0.251
Likelihood of Side-Effects: 20%*	Mean	-1.0219	0.1064	-1.2305	-0.8132	<0.001
	SD	1.1816	0.0699	1.0446	1.3186	<0.001
T_X Effectiveness/Failure: 95%/5%	(ref)					
T_X Effectiveness/Failure: 90%/10%	Mean	-0.1770	0.0879	-0.3492	-0.0048	0.044
	SD	0.5362	0.1732	0.1967	0.8756	0.002
T_X Effectiveness/Failure: 85%/15%	Mean	-0.6337	0.0935	-0.8169	-0.4506	<0.001
	SD	0.2842	0.3465	-0.3950	0.9635	0.412
T_X Effectiveness/Failure: 80%/20%	Mean	-0.8753	0.0706	-1.0136	-0.7370	<0.001
	SD	0.9317	0.0813	0.7724	1.0911	<0.001
Cost: €10	(ref)					
Cost: €25	Mean	-0.2265	0.0455	-0.3157	-0.1374	<0.001
	SD	-0.0354	0.1158	-0.2624	0.1916	0.760
Cost: €40	Mean	-0.8017	0.0532	-0.9059	-0.6975	<0.001
	SD	-0.6719	0.0700	-0.8092	-0.5347	<0.001
Cost: €100	Mean	-2.6369	0.1189	-2.8700	-2.4039	<0.001
	SD	2.7628	0.1153	2.5368	2.9887	<0.001
Interaction effects (Frame*Parameter)[*]						
Effectiveness* Duration: 7 days	Mean	-0.1595	0.0932	-0.3421	0.0232	0.087
Combination* Duration: 7 days	Mean	-0.2188	0.0925	-0.4001	-0.0375	0.018
Effectiveness* Duration: 14 days	Mean	0.0266	0.0996	-0.1687	0.2219	0.789

(cont'd)

		random effects model				
				95% CI		
Attribute: Level		**β-coeff**	**SE**	**Lower**	**Upper**	*P*
Combination* Duration: 14 days	Mean	-0.0783	0.0982	-0.2707	0.1141	0.425
Effectiveness*Side-Effects: 5%	Mean	-0.2160	0.1160	-0.4434	0.0113	0.063
Combination*Side-Effects: 5%	Mean	0.0361	0.1216	-0.2023	0.2746	0.766
Effectiveness*Side-Effects: 10%	Mean	-0.6317	0.1144	-0.8559	-0.4074	<0.001
Combination*Side-Effects: 10%	Mean	-0.3367	0.1145	-0.5611	-0.1123	0.003
Effectiveness*Side-Effects: 20%	Mean	-0.6517	0.1518	-0.9493	-0.3541	<0.001
Combination*Side-Effects: 20%	Mean	-0.3578	0.1511	-0.6539	-0.0616	0.018
Log-likelihood		-8967.0286				

Note: Parameters with significant Random effects are shown with SD; *β Coefficients are for Arm 1 (Failure); T_X: Treatment,* All interaction effects were included in the final MIXL model as fixed effects so no SDs are shown,

The different arms had comparable IVS patterns when standardizing on the most important attribute per arm. Costs and *Contribution to Antibiotic Resistance* were found to have the highest IVS with the individual effectiveness or failure frames valuing *Contribution to Antibiotic Resistance* slightly higher and those who saw the combination frame valuing costs slightly higher. After these, all three arms had similar attribute valuations with *Likelihood of Side-Effects*, *Treatment Outcome*, and Treatment Duration being the third, fourth, and fifth most important attributes. The primary difference was that the failure frame had very similar IVS for *Likelihood of Side-Effects* and *Treatment Outcome* while the combination and effectiveness frame had more pronounced differences between the attribute valuations. When the IVS were standardized on the attribute of *Treatment Outcome*, the differences in scale parameter become more visible as the scale of the Failure framing is significantly smaller than that of Effectiveness of Combined frames. The IVSs are visualized in Figure 2.

Choice predictions

Choice predictions based on individual parameter estimates were similar between arms. The most preferred antibiotic was predicted to be chosen by 92.3-95.2% of participants. The least preferred antibiotic option was predicted to be chosen by 0.3- 0.5% of participants when compared to the most preferred antibiotic. The Amoxicillin was predicted to be chosen by 4.2-7.3% of participants when compared to the most preferred antibiotic. See Table 5 for predicted choice predictions by frame.

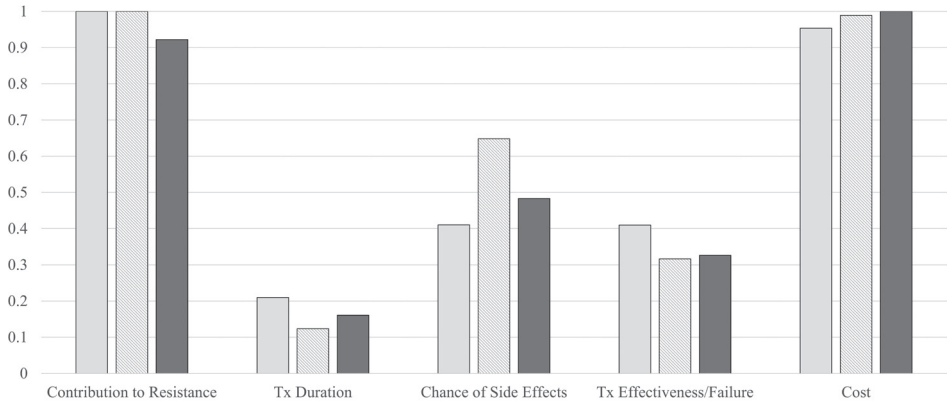


Figure 2 Attribute IVS when standardized on most important attributed (left) or on Tx outcome (right).
Notes: Key: Light grey: failure framing; striped: effectiveness framing; dark grey: combined frames.
Abbreviations: IVS, importance value scores; Tx, treatment.

Table 5. Choice Predictions by arm (Individual Uptake Probability %):

Framing	Most preferred antibiotic[a]	Least preferred antibiotic[b]	Amoxicillin proxy[c]
Failure Rate	95.23	0.53	4.24
Effectiveness	92.39	0.27	7.34
Combined	93.32	0.27	6.42

Note: a: Most preferred antibiotic defined as low contribution to antibiotic resistance, a treatment duration of 3 days (7 days for failure framing), 1% of users experiencing a side effect, an 95%/5% effectiveness/failure rate (90%/10% for combination framing), costing €10; b: Least preferred defines as high contribution to antibiotic resistance, a treatment duration of 14 days, 20% of users experiencing a side effect, an 80%/20% effectiveness/failure rate, costing €100; c: Amoxicillin defined as high contribution to antibiotic resistance, a treatment duration of 7 days, 10% of users experiencing a side effect, an 85%/15% effectiveness/failure rate, costing €10;

DISCUSSION

This study was the first that we know of that used a randomized design to compare the impact of framing when an attribute in a health oriented DCE is framed as positive, negative, or a combination of different frames. This study showed that the impact of framing should not be evaluated simply in regards to the attribute in question. Rather, the way that an attribute is framed impacts the concurrent valuation of other attributes in a DCE, altering the utility of the alternative as a whole which is similar to previous findings.[24] These findings reflect the foundational theory underlying DCEs that the coefficients represent not only the valuation of the individual attributes, but the valuation of these attributes in relation to one another and the decision context in which they are being valued.[7]

In our study, those who saw the treatment outcome framed as effectiveness were more concerned with the likelihood of side-effects in their decision making than those who saw

treatment outcome framed in terms of failure. One possible explanation for this is a negativity bias where participants' attention is drawn away from positively framed attributes to more negative ones.[50, 51] The potential presence of this bias is reflected in the IVS of the different frames. We found that participants who saw a positive frame (effectiveness) had a greater importance value for the negative attribute of likelihood of side-effects compared to those who saw a negative frame. For the negative framed participants, the importance value of these two negative attributes were essentially the same. A similar explanation could also be that participants were more averse to risk in a positive treatment context than in a negative treatment context. Thus, participants were more concerned with the risk of side-effects when they believed a treatment to be effective than when a treatment was believed to not be effective. This type of risk aversion also reflects previous psychological research.[18] Interestingly, these significant differences occurred without a statistically significant difference in the valuation of the treatment effectiveness attribute itself.

This study builds upon existing research looking at how attribute framing impacts the preferences of participants when elicited in a DCE. The framing effects that we found were smaller than those found in other preference studies.[23-26] The different preference estimates did not translate into largely different uptake rate projections. These comparable choice predictions would likely not lead to drastically different decisions being made by stakeholders who use this type of information to guide decision-making. Thus, the practical impact of this framing effect may be limited.

The question then is should anything be done to account for this effect. Druckman previously argued that presenting an attribute with both frames would serve as a type of baseline, eliminating the effect of presenting only one attribute.[27] This was not supported in our study where the two most important attributes were reversed for the combination frame compared to the individually framed alternatives. Additionally, the combination framing did not seem to present a "middle of the road" solution as the attribute level estimates framing did not fall between the positive and negative frames for any attribute across all levels. This seems to imply that the combination frame is a new frame in and of itself rather than a way to compensate for differences in opposing frames and thus cannot be used to account for a framing effect. Howard and Salkeld proposed reviewing attributes for potential framing issues during the development of the DCE and adjusting for a framing effect in the study design by presenting both frames or randomizing to different presentations of attributes.[26] Our study highlights issues with these recommendations as the presentation of multiple frames resulted in a new framing bias which did not necessarily reflect a middle point between the two frames across all attributes. This was especially evident for the importance value of the two most important attributes. Further, randomization to different frames between arms does not provide the relevant information needed to account for framing biases without increasing the number of responses needed for comparison. Randomization within a survey (presenting alternating

frames between choice tasks) may be a way to see how large this framing effect is, but could not account for it and would require an increase in the number of choice scenarios needed to derive the requisite parameters that the participant would essentially need to complete two DCEs.

All of these suggestions treat framing effects as something to control for. It may be that the way that an attribute is framed fundamentally changes the decision context in a way that cannot and should not be corrected for. The outcomes resulting from stated preferences in these different contexts represent the "true" preferences of the participant in that context that they are presented with. Best practice guidelines for conducting a DCE say that attributes and levels should be developed through a qualitative process including contributions from representatives of all stakeholders who participant in or use the outcomes from preference studies.[8, 13, 14] As this study has shown, an important aspect of this process should also be to identify attributes which may be sensitive to framing effects and see how the different frames impact their relevance to these stakeholders. These do not necessarily need to be used to correct for this effect, but rather should be used to understand the preferences in varied contexts which reflect the variance seen in consultation rooms and information presented to patients. A potential area for further research would be a qualitative study to assess why the different frames had the impact that they did. Understanding the working mechanism would give a better understanding of how the different frames impact valuation. This would also help to determine what actions, if any, should be taken in situations where a framing effect may occur. Without this qualitative information, we can only make conjectures about why different valuations were found in the different frames.

A strength of this study is that the randomization process resulted in three groups that were comparable to each other and the general Swedish population in regard to demographic and psychosocial aspects. This limits the chance that differences between the groups impacted the outcomes that were found and supports the generalizability of these findings to the general population. Further, the preference scenario used a medication which many people were already familiar with supporting the validity of the findings as participants had personal experiences to draw from. One limitation, however, is that relative lack of information on participants with a lower education levels or lower health literacy. While the demographic make-up was similar to the general Swedish population, the framing effect in these groups may be more pronounced. Another limitation is that the choice scenario was not necessarily realistic as patients are not often given different antibiotic treatment options by their treating physicians. Thus, the decision scenario was almost entirely hypothetical. Different framing effect outcomes may be found when samples consist of patients currently in treatment with decision contexts relevant to this care. One aspect that was not checked between the different arms was whether participants reinterpreted the outcomes to view them in different frames. That is, did a respondent who saw a treatment profile with 90% effectiveness see this in terms of the 90% effectiveness or as a 10% failure. While this is hypothetically possible, previous research has not shown that

patients reinterpret frames in this way and we would have likely seen higher concordance between the combination arm and a single framed arm if this were the case. Future studies should consider qualitative research to see whether patients reinterpret these frames. Similarly, as this was a hypothetical situation it is not possible to know whether the participants would truly make the same choice if presented with these options in their actual care.

CONCLUSION

This study reaffirmed the impact that framing can have on preference outcomes. In a first of its kind design, three frames (including a combined frame) were used to evaluate the impact of framing on an attribute describing treatment effectiveness. While framing effects were found, the practical implications and interpretation of preferences outcomes was not drastically changed because of this bias. A theoretical discussion is needed to address how to move forward in light of these results. Specifically, addressing the question of whether framing effects should be controlled for, or simply better understood through improved qualitative research.

REFERENCES

1. De Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Economics. 2012;21(2):145-172.
2. Clark MD, Determann D, Petrou S, Moro D, De Bekker-Grob EW. Discrete Choice Experiments in Health Economics: A Review of the Literature. PharmacoEconomics. 2014;32(9):883-902.
3. Soekhai V, De Bekker-Grob EW, Ellis AR, Vass CM. Discrete Choice Experiments in Health Economics: Past, Present and Future. PharmacoEconomics. 2019;37(2):201-226.
4. Lancaster KJ. A New Approach to Consumer Theory. Journal of Political Economy. 1966;74(2):132-157.
5. Manski CF. The structure of random utility models. Theory and Decision. 1977;8(3):229-254.
6. McFadden D. Econometric models of probabilistic choice. Structural analysis of discrete data with econometric applications. 1981;198272.
7. Ryan M, Amaya-Amaya M. 'Threats' to and hopes for estimating benefits. Health Economics. 2005;14(6):609-619.
8. Hollin IL, Craig BM, Coast J, et al. Reporting Formative Qualitative Research to Support the Development of Quantitative Preference Study Protocols and Corresponding Survey Instruments: Guidelines for Authors and Reviewers. The Patient - Patient-Centered Outcomes Research. 2020;13(1):121-136.
9. Bridges JFP, Hauber AB, Marshall D, et al. Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health. 2011;14(4):403-413.
10. Lancsar E, Louviere J. Conducting Discrete Choice Experiments to Inform Healthcare Decision Making. PharmacoEconomics. 2008;26(8):661-677.
11. Helter TM, Boehler CEH. Developing attributes for discrete choice experiments in health: a systematic literature review and case study of alcohol misuse interventions. Journal of Substance Use. 2016;21(6):662-668.
12. Coast J, Al-Janabi H, Sutton EJ, et al. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. Health Economics. 2012;21(6):730-741.
13. Coast J, Horrocks S. Developing attributes and levels for discrete choice experiments using qualitative methods. Journal of Health Services Research & Policy. 2007;12(1):25-30.
14. Kløjgaard ME, Bech M, Søgaard R. Designing a Stated Choice Experiment: The Value of a Qualitative Process. Journal of Choice Modelling. 2012;5(2):1-18.
15. Administration UDoHaHRFaD. Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims In. Silver Spring, MD2019.
16. Lichtenstein S, Slovic P. The construction of preference. Cambridge University Press; 2006.
17. Slovic P. The construction of preference. American psychologist. 1995;50(5):364.
18. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981;211(4481):453-458.
19. Adrian Edwards GEJ. Presenting Risk Information A Review of the Effects of Framing and other Manipulations on Patient Outcomes. Journal of Health Communication. 2001;6(1):61-82.
20. Visschers VHM, Meertens RM, Passchier WWF, De Vries NNK. Probability Information in Risk Communication: A Review of the Research Literature. Risk Analysis. 2009;29(2):267-287.

21. Gong J, Zhang Y, Yang Z, Huang Y, Feng J, Zhang W. The framing effect in medical decision-making: a review of the literature. Psychology, Health & Medicine. 2013;18(6):645-653.

22. Peng J, Jiang Y, Miao D, Li R, Xiao W. Framing effects in medical situations: Distinctions of attribute, goal and risky choice frames. Journal of International Medical Research. 2013;41(3):771-776.

23. McNeil BJ, Pauker SG, Sox HC, Tversky A. On the Elicitation of Preferences for Alternative Therapies. New England Journal of Medicine. 1982;306(21):1259-1262.

24. Veldwijk J, Essers BAB, Lambooij MS, Dirksen CD, Smit HA, de Wit GA. Survival or Mortality: Does Risk Attribute Framing Influence Decision-Making Behavior in a Discrete Choice Experiment? Value in Health. 2016;19(2):202-209.

25. Grisolía JM, Longo A, Hutchinson G, Kee F. Comparing mortality risk reduction, life expectancy gains, and probability of achieving full life span, as alternatives for presenting CVD mortality risk reduction: A discrete choice study of framing risk and health behaviour change. Social Science & Medicine. 2018;211:164-174.

26. Howard K, Salkeld G. Does Attribute Framing in Discrete Choice Experiments Influence Willingness to Pay? Results from a Discrete Choice Experiment in Screening for Colorectal Cancer. Value in Health. 2009;12(2):354-363.

27. Druckman JN. Evaluating framing effects. Journal of Economic Psychology. 2001;22(1):91-101.

28. Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel PA. Effect of Framing as Gain versus Loss on Understanding and Hypothetical Treatment Choices: Survival and Mortality Curves. Medical Decision Making. 2002;22(1):76-83.

29. Ancillotti M, Stefan Eriksson S, Andersson DI, Godskesen T, Nihlén Fahlquist J, Veldwijk J. Preferences regarding antibiotic treatment and the role of antibiotic resistance: a discrete choice experiment. International Journal of Antimicrobial Agents. 2020:106198.

30. Shallcross L, Beckley N, Rait G, Hayward A, Petersen I. Antibiotic prescribing frequency amongst patients in primary care: a cohort study using electronic health records. Journal of Antimicrobial Chemotherapy. 2017;72(6):1818-1824.

31. Poole NM, Shapiro DJ, Fleming-Dutra KE, Hicks LA, Hersh AL, Kronman MP. Antibiotic Prescribing for Children in United States Emergency Departments: 2009–2014. Pediatrics. 2019;143(2):e20181056.

32. Fleming-Dutra KE, Hersh AL, Shapiro DJ, et al. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. JAMA. 2016;315(17):1864.

33. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. The Patient. 2015;8(5):373-384.

34. Sweden TPHAo. Future costs of antibiotic resistance: Final reporting of Government commission on direct and indirect costs and consequences of antibiotic resistance in Swedish health care. 2018.

35. Swedres-Svarm. Consumption of antibiotics and occurrence of antibiotic resistance in Sweden. The Public Health Agency of Sweden, National Veterinary Institute;2018.

36. Veldwijk J, Lambooij MS, De Bekker-Grob EW, Smit HA, De Wit GA. The Effect of Including an Opt-Out Option in Discrete Choice Experiments. PLoS ONE. 2014;9(11):e111805.

37. Wångdahl JM, Mårtensson LI. The Communicative and Critical Health Literacy Scale – Swedish Version. Scandinavian Journal of Public Health. 2013;42(1):25-31.

38. McNaughton CD, Cavanaugh KL, Kripalani S, Rothman RL, Wallston KA. Validation of a Short, 3-Item Version of the Subjective Numeracy Scale. Medical Decision Making. 2015;35(8):932-936.

39. Swait J, Louviere J. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. Journal of Marketing Research. 1993;30(3):305-314.

40. Bansback N, Hole AR, Mulhern B, Tsuchiya A. Testing a discrete choice experiment including duration to value health states for large descriptive systems: Addressing design and sampling issues. 2014;114:38-48.

41. Hauber AB, González JM, Groothuis-Oudshoorn CGM, et al. Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value in Health. 2016;19(4):300-315.

42. Hensher DA, Greene WH. The Mixed Logit model: The state of practice. Transportation. 2003;30(2):133-176.

43. Train KE. Drawing from Densities. In: Train KE, ed. Discrete Choice Methods with Simulation. 2 ed. Cambridge: Cambridge University Press; 2009:205-236.

44. Gonzalez JM. A Guide to Measuring and Interpreting Attribute Importance. The Patient - Patient-Centered Outcomes Research. 2019;12(3):287-295.

45. Control CfD, Prevention, Control CfD, Prevention. Outpatient antibiotic prescriptions—United States, 2014. Data Sources: Systemic oral antibiotics were extracted from the QuintilesIMS Xponent database QuintilesIMS captured an estimated. 2017;88:1308-1316.

46. de Greeff S, Mouton J, Schoffelen A, Melles D, Mevius D, Natsch S. Consumption of antimicrobial agents and antimicrobial resistance among medically important bacteria in The Netherlands in 2015: NethMap 2016. Dutch Foundation of the Working Party on Antibiotic Policy (SWAB);2016.

47. Gillies M, Ranakusuma A, Hoffmann T, et al. Common harms from amoxicillin: a systematic review and meta-analysis of randomized placebo-controlled trials for any indication. Canadian Medical Association Journal. 2015;187(1):E21-E31.

48. Henry DC, Riffer E, Sokol WN, Chaudry NI, Swanson RN. Randomized Double-Blind Study Comparing 3- and 6-Day Regimens of Azithromycin with a 10-Day Amoxicillin-Clavulanate Regimen for Treatment of Acute Bacterial Sinusitis. Antimicrobial Agents and Chemotherapy. 2003;47(9):2770-2774.

49. Sjöberg D, Johansson O, Grundström U, Dalin S. International price comparison 2019: An analysis of Swedish pharmaceutical prices and volumes relative to 19 other European countries. Dental and Pharmaceutical Benefits Agency;2019.

50. Rozin P, Royzman EB. Negativity Bias, Negativity Dominance, and Contagion. Personality and Social Psychology Review. 2001;5(4):296-320.

51. Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD. Bad is Stronger than Good. Review of General Psychology. 2001;5(4):323-370.

Chapter 6

Discrete choice experiment versus swing-weighting: A head-to-head comparison of diabetic patient preferences for glucose-monitoring devices

Authors: Chiara L Whichello, Ian P Smith, Jorien Veldwijk, G. Ardine de Wit, Maureen PMH Rutten-van Molken, Esther W de Bekker-Grob

Status: Published

Whichello, C., Smith, I., Veldwijk, J., de Wit, G.A., Rutten-van Molken, M.P. and de Bekker-Grob, E.W., 2023. Discrete choice experiment versus swing-weighting: A head-to-head comparison of diabetic patient preferences for glucose-monitoring devices. PLoS One, 18(7), p.e0283926.

ABSTRACT

Introduction: Limited evidence exists for how patient preference elicitation methods compare directly. This study compares a discrete choice experiment (DCE) and swing-weighting (SW) by eliciting preferences for glucose-monitoring devices in a population of diabetes patients.

Methods: A sample of Dutch adults with type 1 or 2 diabetes (n=459) completed an online survey assessing their preferences for glucose-monitoring devices, consisting of both a DCE and a SW exercise. Half the sample completed the DCE first; the other half completed the SW first. For the DCE, the relative importance of the attributes of the devices was determined using a mixed-logit model. For the SW, the relative importance of the attributes was based on ranks and points allocated to the 'swing' from the worst to the best level of the attribute. The preference outcomes and self-reported response burden were directly compared between the two methods.

Results: Participants reported they perceived the DCE to be easier to understand and answer compared to the SW. Both methods revealed that cost and precision of the device were the most important attributes. However, the DCE had a 14.9-fold difference between the most and least important attribute, while the SW had a 1.4-fold difference. The weights derived from the SW were almost evenly distributed between all attributes.

Conclusions: The DCE was better received by participants, and generated larger weight differences between each attribute level, making it the more informative method in our case study. This method comparison provides further evidence of the degree of method suitability and trustworthiness.

INTRODUCTION

The integration of patient preferences into decision-making is becoming progressively more important throughout the medical product lifecycle (MPLC) [1]. Projects such as IMI-PREFER [2] and the MDIC (Medical Device Innovations Consortium) [3] are promoting the importance of patient preference information in benefit-risk assessments, while the National Institute for Health and Care Excellence (NICE) is establishing patient preference research partnerships [4]. There is consensus among industry, regulatory, and health technology assessment (HTA) stakeholders that patient preference information would be beneficial when informing benefit-risk assessments throughout the MPLC [2, 5]. This includes the selection of endpoints in early clinical development, to inform regulatory benefit-risk assessments, and to be submitted alongside reimbursement dossiers for HTA appraisal [6].

Different preference measurement techniques exist for specific decision-making contexts. These contexts reflect situations where this patient preference information have high value, such as when there are multiple, alternative treatments with very different benefit-risk profiles [7]. It is vital that decision-makers and researchers select the most appropriate methods suitable for preference-sensitive contexts. However, there is a lack of guidance in current literature regarding the suitability of different patient preference elicitation methods for different situations [7, 8]. A recent empirical comparison has identified discrete choice experiments (DCE) and swing-weighting (SW) as being among the most promising methods likely to meet decision-makers' needs throughout the MPLC [8].

DCEs are derived from random utility theory (RUT), and assume that a healthcare intervention can be represented by its characteristics (also called attributes) [9, 10]. The relative importance of these attributes can be determined by presenting a series of hypothetical choice tasks, and asking for participants' preferred option. The relative weights for each attribute and attribute-level can be derived statistically [11]. DCE outcomes can be used to answer a number of different research questions including trade-off quantification, the willingness-to-pay for different alternatives, and expected uptake rates [12].

SW determines the relative importance based on the improvement of an attribute from its worst state to its best state [13, 14]. Each attribute is first ranked by participant reflecting the importance of this 'swing' from the worst to best level, then points are assigned to each ranking during what is referred to as 'point allocation' [15] [16]. SW also assumes that a participant's utility can be summarized by an explainable value where an individual is always assumed to select an alternative with a higher utility.

Both of these methods can be used to assess the relative value that different attributes and attribute-levels have for the participant. However, whether different methods lead to the same conclusions, when answering the same research question, is a research topic in need of investigation [16]. Literature comparing DCE and SW is lacking [14, 17], although both

methods are increasingly used in healthcare to empirically evaluate the relative desirability of treatment options or attributes [3, 13]. Rating methods, such as swing-weighting, are often regarded as a simpler approach to eliciting patient preferences since they do not force simultaneous trade-offs between multiple attributes [15]. However, other health economists state that direct pairwise comparisons in a DCE are easier for patients than a direct numerical assessment of relative value present in SW [14]. Therefore, the aim of this study is to compare the performance and results of DCE and SW in a common preference context through empirical research.

METHODS

This study compares the DCE and SW in the context of preferences for glucose-monitoring technologies in diabetes patients. Recent advancements in glucose monitoring technology have led to the introduction or new devices to the consumer market such as continuous glucose monitors (CGMs) and flash glucose monitors (FGMs) [18]. These devices are less invasive and more user-friendly than the more commonly used fingerprick-test, which involves direct testing of the blood by lancing the finger multiple times per day to extract a blood sample. The functions, features, and associated costs of CGMs and FGMs vary greatly between devices resulting in a preference sensitive situation [19]. Therefore, the benefit-risk trade-offs affecting a patient's decision for selecting a glucose-monitoring device deserves closer investigation.

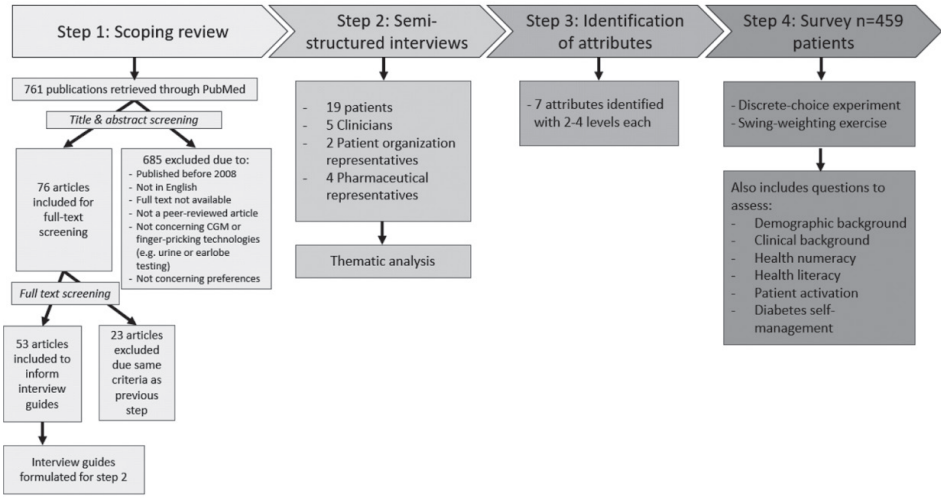


Figure 1: Methodology steps for developing the survey

Attributes and level development

The development of attributes and attribute-levels used to describe the glucose monitoring devices for both the DCE and the SW was conducted in three steps (see Figure 1). In step 1, a scoping literature review was conducted in PubMed to identify relevant attributes of glucose-monitoring devices and develop an interview guide. In step 2, semi-structured interviews were conducted with Type 1 and 2 diabetes patients (n=19), clinicians (n=5), patient organization representatives (n=2), and pharmaceutical industry representatives involved in glucose monitoring device development (n=4) which resulted in an initial list of 12 relevant attributes. In step 3, the list of 12 attributes were rated and reduced according to relevance, completeness, non-redundancy, operationality, and preferential independency by the research team. Subsequently, seven attributes were selected for the DCE and SW (Table 1). The levels used to describe the attributes were based on real-world data [20], representing the most common types of glucose monitoring devices, including CGMs and FGMs [21, 22, 23]. The DCE incorporates all levels, while the methodology of the SW examines only the 'swing' from the lowest level to the highest level. The attributes and levels were presented identically in both methods in order to make accurate comparisons and avoid framing effects. The draft questionnaire was pre-tested during six 'think-aloud' tests, checking for comprehensibility and clarity.

Design

Discrete choice experimental design

NGene 1.0 [24] software was used to develop a Bayesian D-efficient design, consisting of three blocks of 12 choice tasks. Each contained three alternatives (i.e. profiles) with seven attributes of varying levels; two alternatives represented hypothetical glucose-monitoring devices and one represented the fingerpricking test. Participants were given two 'warm-up' DCE choice-tasks before the main exercise in order to ensure comprehension. The questionnaire was tested in a pilot of 99 participants in order to retrieve priors, which informed the design of the final DCE to optimise statistical efficiency.

After the pilot test of the DCE, the DCE design with three alternatives per choice task was substituted by a "best-best" or a so-called 'dual response' DCE [25]. In this task participants were asked which of the two hypothetical device alternatives they would prefer, either 'Device A' or 'Device B'. Then a follow-up task asked if they would prefer the hypothetical device chosen or a standard fingerprick-test (see Appendix I for an example choice task). This design improves data quality by reducing the chance that participants default to the standard opt-out in order to decrease the burden of evaluating the alternatives, while maintaining a realistic decision context in which opting for the fingerprick-test is a reasonable option [26].

Table 1: Attributes and levels for the discrete choice experiment and swing-weighting

Attributes	Level 1	Level 2	Level 3	Level 4
Precision compared to fingerpricking [a]	Less accurate than fingerpricking (higher or lower by 0.6)*	Less accurate than fingerpricking (higher or lower by 0.3)	Accurate as fingerpricking*	---
Average number of fingerpricks per day [b]	4*	2	0*	---
Effort to check [c]	High effort: you need to measure your glucose levels yourself*	Moderate effort: you scan a sensor to check glucose levels	Low effort: glucose levels automatically sent to you*	---
Probability of getting skin irritation or redness [d]	35% chance of skin irritation or redness*	20% chance of skin irritation or redness	5% chance of skin irritation or redness	No chance of skin irritation or redness*
Monthly costs [e]	€250*	€175	€100	€25*
Glucose information [f]	Current Glucose level*	Current Glucose level and arrow	Current Glucose level and a graphic of your level trends over the day*	---
Alarms [g]	No*	Yes*	---	---

* Level included in SW (method only contains highest and lowest levels within attributes)

(a-g) Attribute explanations as presented to patients:

A. Some glucose monitors are more precise than others. Fingerpricking is generally regarded as the most accurate way to measure glucose levels. Measurements from devices that use sensors can be just as accurate, but can also be less accurate than fingerpricking, especially if your glucose levels are very high or very low. For example, if your glucose level is 6 mmol/L and you measure it with a device that is off by 0.6 mmol/L, then this device can say your glucose is anywhere from 5.4 to 6.6 mmol/L

B. This is how many times you would need to do a fingerprick-test each day on an average day. This number could be higher on days when you feel the need to test more often like when you're sick, but we want you to picture an average day. Sometimes, this is your only method of measuring your glucose levels. Or, you might need to do fingerprick-tests to confirm the levels from another device

C. This means how much effort you need to give to check your blood glucose levels. High effort checking means you need to stop what you're doing and concentrate on measuring your levels. You need to wash your hands, get out your device equipment, prick your finger, put blood on a strip, check the results, and then clean everything up. Moderate effort checking means you need to get out a small device and use it to scan the sensor on your body to obtain your glucose levels. Low effort checking means your glucose levels are automatically sent to a device which you can view at any time. This could be a dedicated glucose device, your phone, or a smartwatch. You don't need to do anything to have your blood glucose levels sent through, just look at the device to check.

D. A chance of skin irritation or redness around a sensor means a redness or itchy rash on the skin around or under the sensor. This is similar to having an itchy allergic reaction and can be rather uncomfortable or irritating. The sensor will need to be removed and replaced in a different spot. This skin irritation and redness usually lasts until after the sensor is replaced. Not all sensor have this side effect so chances of getting the side effect can differ per device. If a device gives you a 15% chance, this means that 15 out of a 100 people who get this device experience skin irritation and redness while 85 out of a 100 people do not experience this.

E. This means how much money you need to pay out-of-pocket per month in order to check your blood glucose. Please note that this is money that is not reimbursed by your insurance. This could be money needed to pay for devices, sensors, or strips used.

F. This means how your glucose levels are presented to you. This information could be only your current glucose level (you only see a digital number like 8.3 mmol/L). This could be your current glucose level with an arrow showing how your blood glucose is changing as compared to your previous measurement (increasing, decreasing, stable). Or, it could show your current glucose level with a graphic of your blood glucose levels over the day.

G. Your device will give you a beeping alarm (like a phone notification) any time your blood glucose levels are (getting) too high or too low.

Swing-weighting design

The SW contained two parts. First, participants were asked to rank the seven attributes based on how they would prioritise improving each swing of an attribute-level from its worst to its best state (see Appendix II for an example exercise). The seven attributes were listed randomly for each participant in order to prevent an intrinsic top-down ranking bias. Thereafter, participants were asked to allocate points, from 0 to 100, to each of the swings relative to their first choice which was automatically allocated 100 points [16, 28]. For instruction, participants were informed that if they allocated an attribute 50 points, this indicated they thought improving its state was half as important as their first ranked attribute-level. If participants attempted to allocate more points to a lower-ranked attribute that a higher-ranked attribute, they were presented with a pop-up message drawing attention to this action and ask them to confirm that they wish to proceed with the allocation.

Questionnaire

The questionnaire was online and self-administered. After informed consent, participants received information on the meaning of all the attributes and levels, and then completed demographic questions. All respondents completed both the DCE and SW exercises, but the order was randomised with half of respondents seeing the DCE first and the other half seeing the SW first. Each exercise was followed by debriefing questions related to the ease of understanding the exercise and ease of completing the exercise. Respondents answered on a Likert scale from 1 to 6; 1 being the most difficult and 6 being the easiest. In between the DCE and SW, patients answered questions about their medication and glucose monitoring devices they currently used to control their diabetes, and the frequency of use. At the end of the questionnaire, health literacy and numeracy were assessed using the validated questions of the Shortened Subjective Numeracy Scale (SNS-3) [27] and the Brief Health Literacy Screener (Chew Items) [28].

Members of an online panel who are adult Dutch residents with type 1 or type 2 diabetes were invited to complete the survey. Diabetes diagnosis was self-reported, with no restrictions on type 1 or 2. Information about ethical approval can be found in Appendix III.

Statistical analysis

Discrete choice experiment analysis

The DCE was analysed by combining the outcomes of both best-best tasks into one task comparing all three alternatives (Device A versus Device B versus the fingerprick-test). The outcome of the second best-best task (hypothetical device chosen or fingerprick-test) was used to determine the participant's choice for use in the final model. Observations were analysed in NLOGIT [29] by a latent-class model and a mixed-logit model [30]. Based on model fit,

the mixed-logit was the model best suited to the data and the following utility function was used for the final analyses:

Equations 1-3:

$$V_{Device\ A} = \beta_0 + \beta_1 * precision_{0.3} + \beta_2 * precision_{0.6} + \beta_3 * pricks\ per\ day_{2x} + \beta_4 * effort_{moderate} + \beta_5 * skin\ irritation_{20\%} + \beta_6 * skin\ irritation_{35\%} + \beta_7 * monthly\ costs_{€100} + \beta_8 * monthly\ costs_{€175} + \beta_9 * monthly\ costs_{€250} + \beta_{10} * information_{arrow} + \beta_{11} * information_{trendline} + \beta_{12} * alarms_{none}$$

$$V_{Device\ B} = \beta_1 * precision_{0.3} + \beta_2 * precision_{0.6} + \beta_3 * pricks\ per\ day_{2x} + \beta_4 * effort_{moderate} + \beta_5 * skin\ irritation_{20\%} + \beta_6 * skin\ irritation_{35\%} + \beta_7 * monthly\ costs_{€100} + \beta_8 * monthly\ costs_{€175} + \beta_9 * monthly\ costs_{€250} + \beta_{10} * information_{arrow} + \beta_{11} * information_{trendline} + \beta_{12} * alarms_{none}$$

$$V_{Fingerprick} = \beta_{13}$$

where V represents the total relative utility for an alternative where β1 to β12 are coefficients reflecting the relative importance of each attribute or attribute-level. β13 is an alternative specific constant reflecting the respondents' preference for the fixed alternative of the fingerprick-test over Device B. β0 is a constant term which identifies the respondent's preferences for Device A over Device B, reflecting a left-right bias (i.e. favouring the left option in case the coefficient is significant and has a positive sign). All attributes and attribute levels were included as random parameters, with a normal distribution, accounting for any heterogeneity in the preferences for those attributes. Robust outcomes were generated by applying 14,000 Halton draws.

The mean of the individual uptake probabilities ($\overline{P}$) was determined by estimating the individuals'(i) utility of a device (Vi) compared to the individuals' utility of the fingerprick alternative (Wi), calculating the probability of this choice, and averaging this across all individuals:

Equation 4: $$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{V_i}}{e^{V_i} + e^{W_i}}$$

where i=1 represents the index of summation and n is the total sample size. Effects coding was used, meaning the reference category is coded as -1, which sums the attribute-level coefficients in each category to zero.

Swing-weighting analysis

The SW analysis was conducted by examining each participant's point allocation for each attribute-level improvement relative to the total number of points allocated. Then, the weighted average of each attribute was calculated across the entire participant sample via equation 5:

Equation 5: $$\overline{S}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{x_{i,a}}{\sum_{j=1}^{7} x_{i,j}}$$

where $\overline{S}(a)$ represents the average relative preference score of an attribute (a), xi,a is the points allocated to the attribute by individual i, n is the total number of participants, j is the index of summation for each attribute, and i is the index of summation for each individual.

Comparison of DCE and SW

The two methods were compared in two ways. First, self-reported feedback from participants indicating how easy the method was to understand and answer was used to compare the methods. These results were stratified by the method that was completed first, health literacy, and numeracy. Drop-out rates during the completion of the exercises were also compared as a proxy for participant burden.

Second, a comparison of how important each attribute was reported to be using each method was examined by looking at the proportion of preference for one attribute compared to the summed preferences for all attributes. For the DCE, this involved examining the absolute difference between the best level coefficients and the worst level divided by the sum of all these differences across the attributes. For the SW, one attribute's weight was calculated as a percentage of the total summed attributes' weights (as shown in Equation 5). The relative weights for both methods, reflected as a proportional percentage, were then directly compared.

Sensitivity analysis

A sensitivity analysis was conducted by comparing the weights derived from the point allocation of the SW against weights derived from the ranking portion of the SW calculated using the rank order centroid (ROC) method [15]. The ROC assigns relative weights for each attribute based on the order they were ranked, as defined by Roberts and Goodwin [31]. The proportional ROC weights were also compared against the proportional DCE weights.

In order to determine whether there were significant differences in attribute rankings between the methods, the respondent-level ranking of the attributes in the DCE and the SW were compared using a (generalised) ordered logit model.

RESULTS

Participants' characteristics

A total of 500 participants completed the survey. Participants who completed the survey faster than 70% of the mean response time were excluded, leaving a sample of 459 respondents (Appendix IV). Furthermore, 233 participants completed the DCE first, while 226 participants completed the SW first.

The mean age of all respondents was 51 years old, with a near even split between male and female respondents. Twenty-seven percent of the total sample reported having diabetes Type 1, 69.1% reported having Type 2. Approximately 18.3% of all respondents already used a CGM or FGM, and 54.4% used fingerpricking. About 93.3% reported a "high" or "intermediate" level of education (Appendix IV).

Comparing DCE and SW results

Feedback comparison

Overall, the DCE was reported by participants to be both easier to understand and to complete, compared to the SW (Figure 2). This was true regardless of which method was completed first. Averaged scores for both ease of understanding and ease of answering the DCE were significantly higher (mean=4.71, s.d.= 1.38; mean = 4.60, s.d.=1.36, respectively) than ease of understanding and ease of answering the SW (mean = 3.85, s.d.= 1.68; mean =3.88, s.d.= 1.61, respectively). Both high-literacy and low-literacy participants rated the DCE higher than the SW, as did high-numeracy and low-numeracy participants.

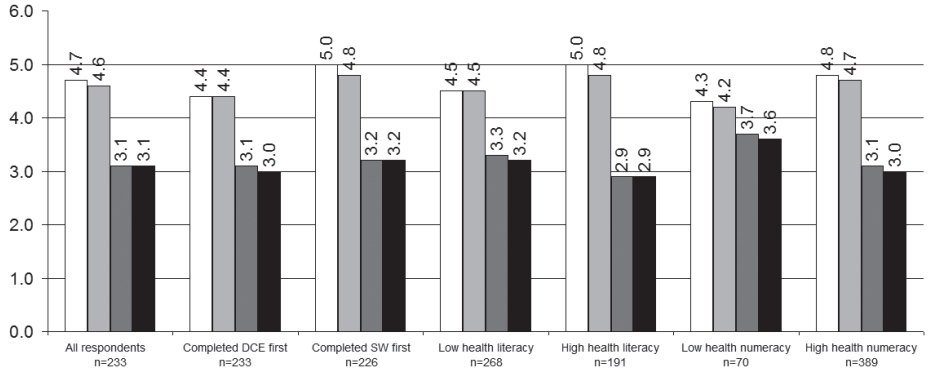


Figure 2: Feedback scores from respondents completing the discrete choice experiment (DCE) and swing-weighting (SW) *White* = Ease of understanding the DCE; *Light grey* = Ease of answering the DCE; *Dark grey* = Ease of understanding the SW; *Black* = Ease of answering the SW; Respondents answered on a scale from 1 to 6, 1 being the most difficult and 6 being the easiest; Low health numeracy scored below 9.83 (the mean) on the SNS-3; High health numeracy scored above 9.83 (the mean) on the SNS-3; Health literacy questions are scored 1-5 with the middle question inversed - Low health literacy identified by a score of ≥3 on any item; High health literacy scored <2 on any item (see Louis et al, 2017 [38]).

Drop-outs were higher during the SW (n=165) than the DCE (n=101), regardless of the order of exercises. Of these, 143 first exercise drop-outs (did not proceed to their second exercise) occurred during the SW, compared to 93 during the DCE.

Discrete choice experiment results

The results of the mixed-logit (Table 2) showed significant estimates for all attribute-levels except for medium precision (0.3 mmol/L), glucose information (information-only), and glucose information with an arrow. Negative coefficients for the attribute-levels indicate that these would not be preferred features in a glucose-monitoring device, relative to the mean

attribute effect. Higher monthly costs were associated with a lower willingness to choose the device. High precision (as accurate as fingerpricking) was strongly preferred over lower precision levels. Respondents generally preferred the fingerprick alternative over either device alternatives presented. However, the model showed significant heterogeneity in respondents' preferences for the constant as well as the other attributes. There also was a slight left-right bias detected.

The predicted uptake rates ranged from 65.9% for the most preferred device (high precision, zero fingerpricks, low effort, low skin irritability, 25 euro, plain information, no alarm) to 10.5% for the least preferred device (low precision, two fingerpricks, moderate effort, high skin irritability, 250 euro, an arrow, an alarm). Individual uptake probabilities did not vary significantly between individuals who saw the DCE first (67.4% for most-preferred device; 10.1% for least-preferred device) and those who saw the SW first (64.3% for most-preferred device; 11.0% least-preferred device).

Swing-weighting results

In general, respondents found cost to be the most important attribute with a mean relative weight of 0.17 (s.d.=0.13), followed by precision (mean=0.16; s.d.=0.12) (see Table 3). The least important attribute was an alarm (mean=0.12; s.d.=0.11). These weights did not vary significantly between individuals who saw the DCE first or the SW first (the difference in mean was <0.02 for all attributes). There was little difference in relative weights given to the seven attributes, with all of the weights being almost evenly distributed across the attributes.

Comparison of weight distribution between the DCE and SW

For the DCE, the proportion of attribute importance is very different for all the attributes (Figure 3). Contrastingly, all attributes in the SW, received between 12-17% of the designated importance. The DCE had a 14.9-fold difference between proportional importance of the most and least important attribute, while the SW had a 1.4-fold difference.

The two attributes with the highest importance were cost and precision, respectively, for both the DCE and SW using point allocation, but the relative weight of costs was much higher in the DCE. For the DCE, the following order of attributes based on their relative importance weight was: skin irritation, fingerpricks, effort, alarms, and glucose information, respectively. For the SW using point allocation, these were fingerpricks, glucose information, effort, skin irritation, and alarms, respectively. The relative weights of all these attributes differed significantly between the two methods.

Table 2: Attribute-level estimates for the discrete choice experiment mixed-logit model

Attribute	Levels		Estimate	p-value	S.E.
Precision compared to fingerpricking	Accurate as fingerpricking (ref)	Mean	0.484	*	0.291
		S.D.	0.762	***	0.006
	0.3	Mean	0.043		0.046
		S.D.	0.087		0.170
	0.6	Mean	-0.527	***	0.068
		S.D.	0.757	***	0.079
Average number of fingerpricks per day	0 times per day (ref)	Mean	0.313	***	0.068
		S.D.	0.532	***	0.003
	2 times per day	Mean	-0.313	***	0.045
		S.D.	0.532	***	0.056
Effort to check	Low (ref)	Mean	0.165	***	0.045
		S.D.	0.231	***	0.003
	Moderate	Mean	-0.165	***	0.033
		S.D.	0.231	***	0.058
Probability of getting skin irritation or redness	5% (ref)	Mean	0.425	***	0.059
		S.D.	0.373	***	0.008
	20%	Mean	-0.091	*	0.050
		S.D.	0.011		0.139
	35%	Mean	-0.334	***	0.056
		S.D.	0.373	***	0.088
Monthly costs	€25 (ref)	Mean	1.728	***	0.096
		S.D.	1.878	***	0.019
	€100	Mean	0.325	***	0.063
		S.D.	0.243		0.162
	€175	Mean	-0.128	*	0.067
		S.D.	0.447	***	0.113
	€250	Mean	-1.925	***	0.139
		S.D.	1.808	***	0.125
Glucose information	Information only (ref)	Mean	-0.133		0.147
		S.D.	0.108	***	0.018
	Arrow	Mean	0.022		0.049
		S.D.	0.055		0.142
	Trendline	Mean	0.111	**	0.049
		S.D.	0.094		0.157
Alarms	Yes (ref)	Mean	0.151	***	0.247
		S.D.	0.348	***	0.003
	No	Mean	-0.151	***	0.036
		S.D.	0.348	***	0.051
Alternative specific constant for fingerprick-test†		Mean	0.949	***	0.287
		S.D.	5.089	***	0.321
Alternative specific constant indicating left-right bias		Mean	0.359	***	.070
		S.D.		**	.103

* indicates p < 0.1 ; ** indicates p < 0.05 ; ** indicates p <0.01; S.D. indicates standard deviation; ref indicates reference level

† This is an alternative specific constant reflecting the respondents' preference for the fixed alternative of the fingerprick-test over Device B. Participants were informed that a fingerprick-test should be done four times a day, requires high effort to check, does not result in skin irritation or redness, will show your glucose levels, doesn't have an alarm and costs €25 per month.

Note: Due to non-linearity of the attributes, all were effects-coded, enabling the direct comparison of the estimates. The sum of the effect coded attributes is zero, and therefore the coefficient of the reference category can be easily calculated and the relative importance of the reference categories of the attributes can be compared with one another, and so that the alternative specific constants have independent interpretation signifying the average utility for that alternative.

Table 3: Swing weighting preference weights, calculated through both point allocation and the rank order centroid (ROC) method

		All respondents (n=459)			**Saw DCE first (n=233)**			**Saw SW first (n=226)**		
Attribute	WCM	Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
Cost	PA	0.17	0.13	0.01	0.18	0.14	0.01	0.16	0.13	0.01
	ROC	0.17	0.13	0.01	0.18	0.13	0.01	0.16	0.13	0.01
Precision	PA	0.16	0.12	0.01	0.17	0.12	0.01	0.16	0.12	0.01
	ROC	0.16	0.11	0.01	0.17	0.11	0.01	0.16	0.11	0.01
Pricks	PA	0.15	0.11	0.01	0.15	0.11	0.01	0.15	0.12	0.01
	ROC	0.18	0.11	0.01	0.19	0.11	0.01	0.17	0.11	0.01
Information	PA	0.14	0.12	0.01	0.15	0.12	0.01	0.14	0.11	0.01
	ROC	0.13	0.11	0.01	0.15	0.12	0.01	0.14	0.11	0.01
Effort	PA	0.13	0.09	>0.00	0.12	0.07	>0.00	0.14	0.11	0.01
	ROC	0.13	0.10	>0.00	0.11	0.09	0.01	0.14	0.11	0.01
Skin	PA	0.12	0.09	>0.00	0.12	0.09	0.01	0.12	0.09	0.01
Irritation	ROC	0.12	0.10	>0.00	0.11	0.10	0.01	0.12	0.10	0.01
Alarms	PA	0.12	0.11	0.01	0.11	0.09	0.01	0.13	0.12	0.01
	ROC	0.11	0.11	0.01	0.11	0.10	0.01	0.12	0.11	0.01

SW = swing-weighting; DCE = discrete choice experiment; WCM = weight calculation method; SD = standard deviation; SE = standard error of mean; PA = point allocation; ROC = rank order centroid.

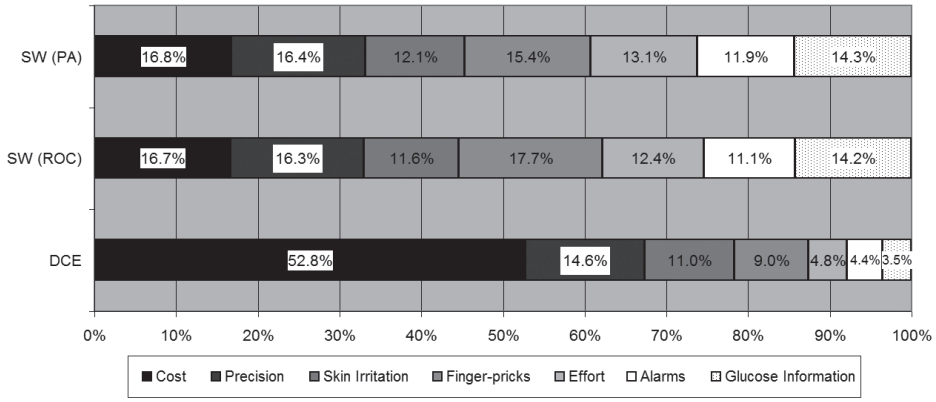


Figure 3: Proportion of attribute importance relative to sum of all attributes' importance (DCE and SW calculated through both ROC and PA Sensitivity analysis

DCE = discrete choice experiment; SW = swing weighting; PA = point allocation; ROC = rank order centroid method

The sensitivity analysis using the ROC instead of the point allocation method to calculate SW, revealed slight differences in the importance of attributes (Table 3 and Figure 3). Fingerprick frequency was the most important attribute (mean=0.18; s.d.=0.11), followed by cost (mean=0.17; s.d.=0.13), with alarm being the least important attribute (mean=0.11; s.d.=0.11). There was a 1.5 fold difference between the proportional importance of the lowest

and highest attribute, compared to the 14.9 -fold difference in the DCE. The ROC method for determining weights still achieved very little difference in the relative weights given to the seven attributes, (Figure 3).

The (generalised) ordered logit model (see Appendix V) also indicated that there were significant differences in the respondent-level rankings of the attributes between the DCE and SW. Fingerpick frequency was more likely to be ranked as the highest attribute in the DCE rather than in the SW and alarms and precision were more likely to be ranked among the bottom-ranked attributes in DCE rather than in the SW. Other attributes showed significant differences in rank order between the two methods as well.

DISCUSSION

Both the DCE and SW point allocation identified that cost was the most important attribute for diabetes patients when selecting a glucose-monitoring device. Preference outcomes in both methods were unaffected by the order in which they were completed. However, the weights derived from the SW were almost evenly distributed regardless of a calculation through point allocation or ROC method. The SW point allocation had a 1.4-fold difference between the most and least important attribute, while the DCE had a 14.9-fold difference. The DCE was better received by participants and obtained more detailed insights for all attribute-levels, making it the preferred method over the SW in this case study.

To the best of our knowledge, this is the first study directly comparing the outcomes of a DCE and a SW task in which the relative weights were able to be compared. Previous research compared the two methods but were unable to directly compare the outcomes [17].

The small difference between the mean attribute weights in the SW warrants further discussion. As the point allocation part of the SW task was a direct rating, participants essentially created the weights themselves thus negating the need for researchers to convert rankings to surrogate weights. Incorporating point allocation into SW is often praised for being a simple way to elicit the relative valuation of the attributes by allowing respondents to directly report this valuation for each attribute ("provid[ing] information on relative importance, whilst remaining relatively uncomplicated"[32]). However, there remains some uncertainty about how the point allocation should be administered. The direct ranking task we used did not force participants to trade-off when allocating points to the different attributes. In this way, there was no cost to valuing one attribute over another like there is in a DCE. Other SW techniques ask participants to designate a proportion of 100 points to each attribute, meaning all attribute weights must add up to 100 [15, 35, 16]. While this results in a trade-off between the attribute valuation this type of task has been found to be less reliable. For our case study, the added complexity of trading off points between seven attributes was deemed to be an unnecessary numerical burden if participants had to monitor the total sum score while awarding points. Additional

complexity may result in random responses, or responses becoming unresponsive to small differences in points. Therefore, in this study, participants could award any points out of 100, and their weights were calculated out of their total sum. One issue that has previously been raised with this type of rating method is the poor discriminatory power resulting from insufficient variability in the point allocation [33]. This tends to muddle the differences in the valuation and was evidenced in our study where over 55% of participants allocated the same number of points to at least two attributes.

The sensitivity analysis of the ROC method was important to identify whether the even distribution of weights was only a product of the point allocation methodology [34]. The same phenomenon occurred with the ROC verifying that it is likely a characteristic of SW analysis itself. Crucially, there were small differences found for the most important attributes. The ROC is often criticised for the extreme weights it places on higher-ranked attributes with minimal difference between the weights of lower-ranked attributes [35]. The findings of this study support the conclusion that point allocation is a more robust weight calculation method than ROC and should be used in future SW studies.

Neither weight-calculation technique of the SW gave as specific an insight as the DCE, which forces choices between pairwise comparisons. DCEs 'decompose' treatment or medical product alternatives into specific attributes describing the element that are most influential to patient decisions. This makes it possible to estimate preferences for more levels per attribute than only the best and worst level which are used in SW [33]. An additional benefit of DCEs are the ability to assess preference heterogeneity using mixed logit models. The value of this method was found in our study results which demonstrated strong preference heterogeneity for most device attributes. Finally, the outcomes of a DCE can be used for more than just relative weights of attributes thus the outcomes of one study can be applied to a wider range of applications [38].

The DCE was better received by participants than the SW, regardless of the order completed, or the level of health literacy or health numeracy reported by the patient. Accuracy of preference measurements is highly dependent on patient understanding. Whether respondents started with DCE or SW did not significantly affect preference outcomes in either method. This suggests that the combination of two methods did not create overwhelming cognitive burden or study fatigue, or that there was not a significant ordering effect regarding the way experimental materials were presented.

Previous literature comparing DCE and SW comprehensibility has been lacking; however, it has been theorised that direct pairwise comparisons in a DCE are easier for patients than a direct numerical assessment of relative value present in SW despite the increased cognitive burden attached to assessing multiple attributes concurrently [14, 39]. Additionally, evidence suggests that rating methods are not observed to be easy cognitive tasks due to the involvement of a predetermined numerical scale, and the complexity of applying it against multiple

attributes [36]. Essentially, it is easier to say which of two attributes is more important, rather than trying to quantify how much more important it is [14]. From the researcher's perspective, the DCE may appear more complex compared to the SW in terms of design and analysis, but respondents view the DCE as the simpler method to understand and complete. A SW is a viable alternative in cases when the number of attributes cannot be feasibly integrated into a DCE [7], or when a sample size is too small for a DCE, such as in the case of rare diseases [8, 38].

Limitations

The length of the survey and number of screening questions could have contributed to the drop-out rate within the panel data or created a higher cognitive burden. Due to confidentially agreements, reminder e-mails could not be sent, and a non-response analysis could not be conducted. It took on average 19.2 minutes to complete the survey, which was faster than expected, and could be a limitation of the study if participants did not spend sufficient time reading the instructions. Participants who completed the survey faster than 70% of the mean response time were excluded (n=41) due to their speed decreasing the chance of them having read all elements of the survey. A sensitivity analysis revealed their exclusion bettered the model fit and decreased left-right bias.

About 93.3% reported a "high" or "intermediate" level of education (defined in Appendix IV) meaning there was an underrepresentation of participants with a low level of education. Approximately 18.3% of all respondents already used a CGM or FGM, and 54.4% used fingerpricking, while 27.2% used neither. Individual uptake probabilities for the most-preferred device compared to fingerpricking varied between CGM/FGM-users and fingerpricking-users (13.7% versus 33.0%, respectively).

The listed order of the attributes remained the same for each choice task of the DCE, with precision listed first and cost last, which means that participants could have ignored attributes in the middle when scanning the choice tasks. However, lexicographic behaviour (i.e. always opting for the best level of one attribute) was very low in the dataset, with sensitivity analysis revealing little difference if these (n=19) participants were removed from the sample. The SW always had its attributes randomised during the ranking exercise.

The feedback for understanding and completing the SW exercise did not distinguish between ranking the attributes and point allocation, so therefore we cannot know which part of the SW the participants found the most difficult. This could have helped understanding whether the point allocation was a valuable addition to the exercise.

The ordered logit was conducted with the ranking information from the SW exercise only. This analysis could not be performed using data from the point allocation, due to 55% of participants allocating an equal number of points to at least two attributes in the point allocation.

Implications for future research

Future research should examine DCE and SW in more head-to-head studies with different populations, different medical products treatments, and different decision contexts in order to examine if the same weight distribution occurs. Variations of the SW point allocation should be examined, reducing the number of attributes, forbidding attributes being allocated the same number of points, or forcing all attributes to add to 100. More studies comparing the point allocation system to the ROC method would also help conclude whether point allocation adds meaningful quantitative insight into preferences, or merely adds cognitive burden to respondents.

CONCLUSIONS

This study compared a DCE with SW by eliciting preference for glucose-monitoring devices in a population of 459 diabetes patients. Both methods identified that cost was the most important attribute when selecting a device, followed by the precision of the device. However, the weights derived from the SW, regardless of a calculation through point allocation or ROC method, were almost evenly distributed between the attributes. The DCE was better received by participants, and generated larger weight differences between each attribute level, making it the more informative method in our case study. This method comparison provides further evidence of the degree of method suitability and trustworthiness of these methods for measuring preferences for decision-making. Further research should compare these methods in different disease areas and decision-contexts.

REFERENCES

[1] M. Smith, T. Hammad, M. Metcalf, B. Levitan, R. Noel, A. Wolka, D. Michaels and F. Johnson, "Patient Engagement at a Tipping Point—The Need for Cultural Change Across Patient, Sponsor, and Regulator Stakeholders: Insights From the DIA Conference," *Therapeutic Innovation & Regulatory Science,* vol. 50, no. 5, pp. 546-53, 2016.

[2] E. de Bekker-Grob, C. Berlin, B. Levitan, K. Raza, K. Christoforidi, I. Cleemput, J. Pelouchova, H. Enzmann, N. Cook and M. Hansson, "Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public–Private Project," *The Patient: Patient Centred Outcomes Research,* vol. 10, no. 3, pp. 263-266, 2017.

[3] Medical Device Innovation Consortium (MDIC), "Medical Device Innovation Consortium (MDIC) Patient Centered Benefit-Risk Project Report: A Framework for Incorporating Information on Patient Preferences regarding Benefit and Risk into Regulatory Assessments of New Medical Technology," 2015.

[4] NICE Pathways, "Managing relapse of myeloma," NICE, 2019.

[5] C. Whichello, E. van Overbeeke, R. Janssens, K. Schölin Bywall, S. Russo, J. Veldwijk, I. Cleemput, J. Juhaeri, B. Levitan, J. Kübler, M. Smith, R. Hermann, M. Englbrecht and et al., "Factors and Situations Affecting the Value of Patient Preference Studies: Semi-Structured Interviews in Europe and the US," *Frontiers in Pharmacology,* vol. 24, no. 1, pp. 57-68, 2019.

[6] J. Bouvy, L. Cowie, R. Lovett, D. Morrison, H. Livingstone and N. Crabb, "Use of Patient Preference Studies in HTA Decision Making: A NICE Perspective," *The Patient,* vol. 13, pp. 145-149, 2020.

[7] E. van Overbeeke, C. Whichello, R. Janssens, J. Veldwijk, I. Cleemput, S. Simoens, J. Juhaeri, B. Levitan, J. Kubler, E. de Bekker-Grob and I. Huys, "Factors and situations influencing the value of patient preference studies along the medical product lifecycle: a literature review," *Drug Discovery Today,* vol. 24, no. 1, pp. 57-68, 2019.

[8] C. Whichello, B. Levitan, J. Juhaeri, V. Patadia, R. DiSantostefano, C. Pinto and E. de Bekker-Grob, "Appraising patient preference methods for decision-making in the medical product lifecycle: An empirical comparison," *BMC Medical Informatics and Decision Making,* p. In Press, 2020.

[9] E. de Bekker-Grob, M. Ryan and K. Gerard, "Discrete choice experiments in health economics: a review of the literature," *Health Economics,* vol. 21, pp. 145-172, 2012.

[10] K. Lancaser, "A new approach to consumer theory," *J Polit Econ,* vol. 74, pp. 132-157, 1966.

[11] V. Soekhai, E. de Bekker-Grob, A. Ellis and C. Vass, "Discrete Choice Experiments in Health Economics: Past, Present and Future," *Pharmacoeconomics,* vol. 37, no. 2, pp. 201-226, 2019.

[12] E. Lancsar and J. Louviere, "Conducting discrete choice experiements to inform healthcare decision-making," *Pharmacoeconomics,* vol. 26, no. 8, pp. 661-77, 2008.

[13] V. Soekhai, C. Whichello, B. Levitan, J. Veldwijk, C. Pinto, J. Gonzalez, B. Donkers, E. van Overbeeke, J. Juhaeri and E. de Bekker-Grob, "Compendium of methods for exploring and eliciting patient preferences in the drug and medical device lifecycle," *Drug Discovery Today,* vol. In Press, 2019.

[14] T. Tervonen, H. Gelhorn, S. Sri Bhashyam and et al, "MCDA swing weighting and discrete choice experiments for elicitation of patient benefit-risk preferences: a critical assessment," *Pharmacoepidemiol Drug Saf,* vol. 26, pp. 1483-1491, 2017.

[15] W. Edwards and F. Barron, "SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement," *Organ Behav Hum,* vol. 60, pp. 306-325, 1994.

[16] B. Levitan, B. Hauber, M. Damiano, R. Jaffe and S. Christopher, "The Ball is in Your Court: Agenda for Research to Advance the Science of Patient Preferences in the Regulatory Review of Medical Devices in the United States," *The Patient,* vol. 10, pp. 531-536, 2017.

[17] M. Rutten-van Molken, F. Leijten, M. Hoedemakers, A. Tsiachristas, N. Verbeek, M. Karimi, R. Bal, A. de Bont, K. Islam, J. Askildsen, T. Czypionka, e. al and S. consortium, "Strengthening the Evidence-Base of Integrated Care for People With Multi-Morbidity in Europe Using Multi-Criteria Decision Analysis (MCDA)," *BMC Health Serv Res,* vol. 18, no. 1, p. 576, 2018.

[18] S. K. Garg and H. K. Akturk, "The Future of Continuous Glucose Monitoring," *Diabetes technology & therapeutics,* vol. 19, no. S3, pp. S1-S2, 2017.

[19] I. Hirsch, D. Armstrong, R. Bergenstal, B. Buckingham, B. Childs, W. Clarke, A. Peters and H. Wolpert, "Clinical Application of Emerging Sensor Technologies in Diabetes Management: Consensus Guidelines for Continuous Glucose Monitoring (CGM)," *Diabetes Technology & Therapeutics,* vol. 10, no. 4, pp. 232-244, 2008.

[20] Food and Drug Administration (FDA), "Blood Glucose Monitoring Test Systems for Prescription Point-of-Care Use," Food and Drug Administration (FDA), 2018.

[21] A. L. Carlson, D. M. Mullen and R. M. Bergenstal, "Clinical use of continuous glucose monitoring in adults with type 2 diabetes," *Diabetes Technology and Therapeutics,* 2017.

[22] L. Heinemann and J. H. DeVries, "Reimbursement for Continuous Glucose Monitoring," *Diabetes Technology and Therapeutics,* 2016.

[23] N. A. Allen, J. A. Fain, B. Braun and S. R. Chipkin, "Continuous glucose monitoring in non-insulin-using individuals with type 2 diabetes: Acceptability, feasibility, and teaching opportunities," *Diabetes Technology and Therapeutics,* 2009.

[24] ChoiceMetrics, *Ngene 1.1.1,* User Manual & Reference Guide, Australia, 2012.

[25] P. Ghijben, E. Lancsar and S. Zavarsek, "Preferences for Oral Anticoagulants in Atrial Fibrillation: a Best–Best Discrete Choice Experiment," *PharmacoEconomics,* vol. 32, no. 11, pp. 1115-1127, 2014.

[26] J. Veldwijk, M. Lambooij, E. de Bekker-Grob, H. Smit and A. de Wit, "The Effect of Including an Opt-Out Option in Discrete Choice Experiments," *PLoS One,* vol. 9, no. 11, p. e111805, 2014.

[27] C. McNaughton, K. Cavanaugh, S. Kripalani, R. Rothman and K. Wallston, "alidation of a Short, 3-Item Version of the Subjective Numeracy Scale," *Medical Decision Making,* vol. 35, no. 8, pp. 932-936, 2015.

[28] L. Chew, K. Bradley and E. Boyko, "Brief questions to identify patients with inadequate health literacy," *Family Medicine,* vol. 36, no. 8, pp. 588-594, 2004.

[29] Econometric Software, Inc. , "Nlogit 6," Econometric Software, Inc., Plainview, N.Y., U.S..

[30] J. Louviere, D. Hensher and J. Swait, Stated choice methods: analysis and application, Cambridge: Cambridge University Press, 2000.

[31] R. Roberts and P. Goodwin, "Weight approximations in multi-attribute decision models," *Journal of Multi-Criteria Decision Analysis,* vol. 11, no. 6, pp. 291-303, 2002.

[32] Myeloma UK, "Measuring patient preferences: An exploratory study to determine how patient preferences data could be used in health technology assessment (HTA)," Myeloma UK, 2019.

[33] A. Mühlbacher, P. Zweifel, A. Kaczynski and F. Johnson, "Experimental measurement of preferences in health care using best-worst scaling (BWS): theoretical and statistical issues," *Health Economics Review,* vol. 6, no. 5, pp. 1-12, 2016.

[34] M. Barfod and S. Leleur, Multi-criteria decision analysis for use in transport decision making (2nd edition), DTU Lyngby: DTU Transport, 2014.

[35] M. Danielson and L. Ekenberg, "Rank Ordering Methods for Multi-criteria Decisions," in *Group Decision and Negotiation. A Process-Oriented View. GDN 2014. Lecture Notes in Business Information Processing, vol 180.*, K. G. H. J. Zaraté P., Ed., Cham, Springer, 2014, pp. 128-135.

[36] J. Krosnick and L. Fabrigar, Designing rating scales for effective measurement in surveys. Survey measurement and process quality, Hoboken, NJ: John Wiley & Sons, Inc. , 1997.

[37] A. Louis, V. Arora, M. Matthiesen, D. Meltzer and V. Press, "Screening Hospitalized Patients for Low Health Literacy—Beyond the REALM of Possibility?," *Health Educ Behav,* vol. 44, no. 3, pp. 360-364, 2016.

[38] World Health Organization, "Global status report on non-communicable diseases," World Health Organization, Geneva, 2014.

[39] W. Ong, S. Chua and C. Ng, "Barriers and facilitators to self-monitoring of blood glucose in people with type 2 diabetes using insulin: a qualitative study," *Patient Pref Adherence,* vol. 8, pp. 237-246, 2014.

[40] L. Heinemann and et al, "Reimbursement for continuous glucose monitoring: A european view," *Journal of Diabetes Science Technology,* vol. 6, no. 6, pp. 1498-1502, 2012.

[41] J. Hibbard, J. Stockard, E. Mahoney and M. Tusler, "Development of the Patient Activation Measure (PAM): Conceptualizing and Measuring Activation in Patients and Consumers," *Health Serv Res,* vol. 39, no. 4, pp. 1005-1026, 2004.

[42] A. Schmitt, A. Gahr, N. Hermanns, B. Kulzer, J. Huber and T. Haak, "The diabetes self-management questionnaire (DSMQ): development and evaluation of an instrument to assess diabetes self-care activities associated with glycaemic control," *Health Qual Life Outcomes,* vol. 11, p. 138, 2013.

[43] M. Korytkowski, D. Bell, C. Jacobsen and R. Suwannasari, "A Multicenter, Randomized, Open-Label, Comparative,Two-Period Crossover Trial of Preference, Efficacy, and Safety Profiles of a Prefilled, Disposable Pen and Conventional Vial/Syringe for Insulin Injection in Patients with Type 1 or 2 Diabetes Mellitus," *Clinical Therapeutics,* vol. 25, no. 11, 2003.

[44] G. Brown, M. Brown, S. Sharma, H. Brown, M. Gozum and P. Denton, "Quality of life associated with diabetes mellitus in an adult population," *Journal of Diabetes and its COmplications,* vol. 14, no. 1, pp. 18-24, 2000.

[45] J. Evans, R. Newton, D. Ruta, T. MacDonald and A. Morris, "ocio-economic status, obesity and prevalence of Type 1 and Type 2 diabetes mellitus," *Diabetic Medicine,* vol. 17, no. 6, pp. 478-80, 2000.

[46] C. Terwee, S. Bot, M. de Boer, D. van der Windt, D. Knol and J. Dekker, "Quality criteria were proposed for measurement properties of health status questionnaires," *J Clin Epidemiol,* vol. 60, no. 1, pp. 34-42, 2007.

SUPPLEMENTAL MATERIAL

Appendix I: Example choice task

Imagine that your doctor told you to check your blood glucose levels at least four times per day. To do this, the doctor offers you different hypothetical devices to choose from.

	Device A		Device B
Precision compared to fingerpricking	Less accurate than fingerpricking (higher or lower by 0.3)		Less accurate than fingerpricking (higher or lower by 0.6)
Average number of fingerpricks per day	0		0
Effort to check	Low effort		Moderate effort
Probability of getting skin irritation or redness	5% chance of skin irritation or redness (5 out of 100)		35% chance of skin irritation or redness (35 out of 100)
Glucose information	Current Glucose level		Current Glucose level and arrow
Alarms	Yes		No
Monthly costs	€25		€175
I prefer:	□		□

If you have to choose between the device you have chosen above and the traditional fingerprick-test to check your glucose levels, which one would you prefer? (Please note that a fingerprick-test should be done four times a day, requires high effort to check, does not result in skin irritation or redness, will show your glucose levels, doesn't have an alarm and costs €25 per month).

Select only one answer

□ **I prefer the device I have selected above**	□ **I prefer the fingerprick-test**

Appendix II: Swing weighting

Part 1: Ranking

If you could improve one characteristic of a glucose monitor from being the worst possibility to the best possibility, which would you improve?

Click on the characteristic that you would like to improve first.

Click on the characteristic that you would want to improve next. Continue until all the characteristics have disappeared.

Note: You can hover your mouse over each characteristic to learn more about it

Which would you choose first?

Glucose information: ***Current glucose level ➔ Current glucose level and a graphic of your level trends over the day***	□
Effort to check: ***High effort (you need to measure your glucose levels yourself) ➔ Low effort (glucose levels automatically sent to you)***	□
Probability of getting skin irritation or redness: ***35% chance (35 out of 100) ➔ No chance***	□
Alarms: ***No ➔ Yes***	□
Average number of fingerpricks per day: ***4 daily ➔ 0 daily***	□
Out of pocket cost per month: ***€250➔ €25***	□
Precision compared to fingerpricking: ***Less accurate than fingerpricking (higher or lower by 0.6) ➔ Accurate as fingerpricking***	□

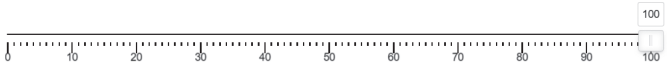
Restart selection

Part 2: Point allocation

Imagine that improving [participant's first choice] is worth 100 points. Please give points to the other characteristics based on how important you think improving them would be compared to improving [participant's first choice].

You can give any number of points to each improvement from 0 (not at all important) to 100 (just as important as [participant's first choice]). For example, if you give 50 points to an improvement, it means that you think it is half as important as [participant's first choice] because you gave it half as many points.

[Participant's 1st choice]



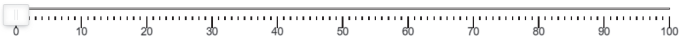
[Participant's 2nd choice]

[Participant's 3rd choice]

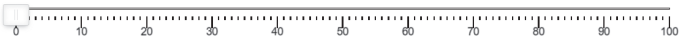
[Participant's 4th choice]

[Participant's 5th choice]

[Participant's 6th choice]



[Participant's 7th choice]



Appendix III: Ethical Approval

On 17 December 2019, this study (Reference number WAG/mb/19/045208) was granted approval by the Medical Research Ethics Committee, UMC Utrecht and confirmed that the Medical Research Involving Human Subjects Act 1998 (Wet Medisch-Wetenschappelijk Onderzoek Met Mensen (WMO)) does not apply to the study because (1) it does not concern medical scientific research and (2) participants are not subject to procedure or are required to follow rules of behaviour. The study was conducted according to the principles of the Declaration of Helsinki.

Appendix IV: Respondent characteristics (n=459)

Characteristics	All patients n=459
Age in years (mean ± sd)	51.0 ± 17.5
Sex (n, %)	
Females	233 (50.8)
Males	225 (49.0)
Glucose monitor currently used	
CGM or FGM	84 (18.3)
Fingerprick-testing only	250 (54.4)
None	125 (27.2)
Type of diabetes (n, %)	
Type 1	124 (27.0)
Type 2	317 (69.1)
Other	18 (3.9)
Educational level [a] (n, %)	
High	184 (40.1)
Intermediate	244 (53.2)
Low	31 (6.8)

CGM= continuous glucose monitor; FGM = 'Flash' glucose monitor

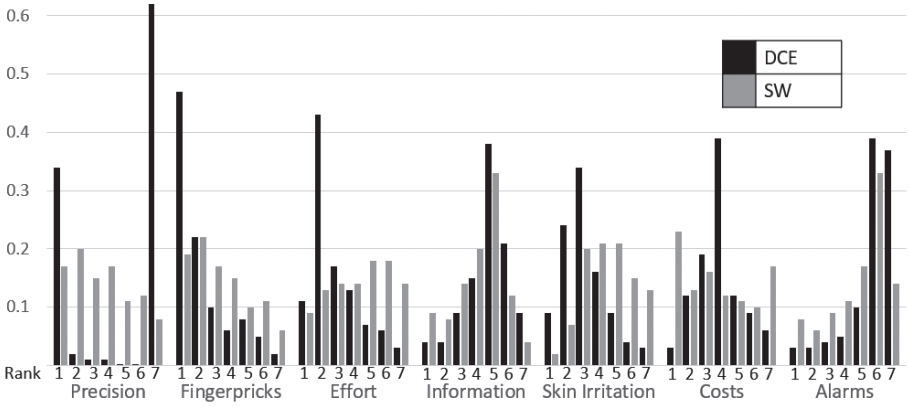
(a) High represents Bachelors, masters, or higher degree; Intermediate represents general secondary education, vocational secondary education, or gymnasium; Low represents lower secondary (high school), primary school, or no education.

Appendix V: Comparison of the probability that an attribute has a certain rank between DCE and SW (Table)

		Attributes						
Probability of Attribute Having Certain Ranks 1-7	**Method**	Precision compared to finger-pricking†	Average number of finger-pricks per day†	Effort to check†	Glucose information‡	Probability of getting skin irritation or redness‡	Monthly costs†	Alarms‡
Rank 1 (highest)	DCE	0.34	0.47	0.11	0.04	0.09	0.03	0.03
	SW	0.17	0.19	0.09	0.09	0.02	0.23	0.08
Rank 2	DCE	0.02	0.22	0.43	0.04	0.24	0.12	0.03
	SW	0.20	0.22	0.13	0.08	0.07	0.13	0.06
Rank 3	DCE	0.01	0.10	0.17	0.09	0.34	0.19	0.04
	SW	0.15	0.17	0.14	0.14	0.20	0.16	0.09
Rank 4	DCE	0.01	0.06	0.13	0.15	0.16	0.39	0.05
	SW	0.17	0.15	0.14	0.20	0.21	0.12	0.11
Rank 5	DCE	0.002	0.08	0.07	0.38	0.09	0.12	0.10
	SW	0.11	0.10	0.18	0.33	0.21	0.11	0.17
Rank 6	DCE	0.002	0.05	0.06	0.21	0.04	0.09	0.39
	SW	0.12	0.11	0.18	0.12	0.15	0.10	0.33
Rank 7 (lowest)	DCE	0.62	0.02	0.03	0.09	0.03	0.06	0.37
	SW	0.08	0.06	0.14	0.04	0.13	0.17	0.14
Number of observations		918	918	918	918	918	918	918
Wald chi2		444.98	89.88	156.02	31.78	127.53	156.97	84.40
p-value		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0014	<0.0001

DCE = Discrete choice experiment; SW = Swing weighting.
†Based on generalized ordered logit analyses; ‡Based on ordered logit analyses because the generalised ordered logit did not converge

Appendix VI: Comparison of the probability that an attribute has a certain rank between DCE and SW (Figure)



Chapter 7

Comparing Discrete Choice Experiment with Swing Weighting to Estimate Attribute Relative Importance: A Case Study in Lung Cancer Patient Preferences

Authors: J. Veldwijk*, I.P. Smith*, S. Oliveri, S. Petrocchi, M.Y. Smith, L. Lanzoni, R. Janssens, I. Huys, G.A. de Wit**, C.G.M Groothuis-Oudshoorn**

*Shared first author
**Shared last author
Status: Published

Veldwijk J, Smith IP, Oliveri S, et al. Comparing Discrete Choice Experiment with Swing Weighting to Estimate Attribute Relative Importance: A Case Study in Lung Cancer Patient Preferences. Medical Decision Making. 2024;0(0). doi:10.1177/0272989X231222421

ABSTRACT

Introduction: Discrete choice experiments (DCE) are common patient preference elicitation methods for attribute relative valuations, but can be complex and costly to administer. Simpler to administer methods which measure relative valuations exist (Swing Weighting with Direct Rating (SW-DR)), but there is little empirical evidence comparing the two. This study aimed to directly compare attribute relative importance when elicited using a DCE and SW-DR.

Methods: 307 non-small cell lung cancer patients in Italy and Belgium completed an online survey assessing preferences for cancer treatment using DCE and SW-DR. Relative importance of the attributes was determined using a Random Parameter Logit model for the DCE and Rank Order Centroid Method for SW-DR. Differences in relative importance ranking and weights between the methods were assessed using Cohens's weighted kappa and Dirichlet regression. Feedback on ease of understanding and answering the two tasks was also collected.

Results: Chance of 5-year survival and risk of extreme tiredness were the most important attributes in both countries regardless of method. The relative ranking and weight of the other attributes significantly differed between DCE and SW-DR; DCEs had greater weight differences between the most and least important attributes. Most respondents found both tasks very easy or easy to understand and answer.

Conclusion: Greater differences in the DCE-derived relative importance valuations may better reflect the type of forced trade-offs stakeholders are interested in when evaluating medical products. Further research comparing the two methods in other choice contexts will help guide researchers in identifying best methods for relative valuation.

INTRODUCTION

As healthcare systems evolve towards more patient-centered drug development, evaluation, and care, there has been an increased interest in using patient preferences to support decision making when developing and evaluating these medical products[1-3]. Patient preference assessments measure what patients value in their healthcare and can be used to compare different aspects of care and trade-offs patients find acceptable [4, 5]. Patient preferences can be explored using a variety of qualitative and quantitative methods [6]. In view of many existing methods to assess patients' preferences, stakeholders have identified a need to compare methods in order to help guide method selection for use in patient preference studies [7].

One frequently used method to elicit and quantify patient preferences is a Discrete Choice Experiment (DCE)[8]. DCEs are based on the Random Utility Theory (RUT) and require respondents to answer several choice tasks in which they are presented with multiple alternatives representing different healthcare options. The alternatives are described using a set of attributes with varying levels [9-11]. From these alternatives respondents choose the option with the highest personal utility [12-15]. Based on the choices respondents make, the impact each attribute has on the utility is estimated and relative importance of the included attributes can be inferred from these estimates [12, 16, 17]. While the validity of DCE findings is well supported [4, 18], DCEs can be complex to administer as they require formal experimental designs [19], complex statistical modelling techniques [12], can be cognitively burdensome to respondents [20, 21], and require relatively large sample sizes with associated higher costs [22, 23].

Another method for preference elicitation and estimation of relative importance ranking and weighting is swing weighting (SW)[8, 24, 25]. In SW tasks, respondents are presented with a list of attributes used to define a healthcare treatment option. Each attribute on the list shows the 'swing' from what patients or researchers find to be its worst level to its best level. The participant ranks these swings based on how important improving that attribute is to them. SW tasks are often followed by a point allocation (PA) or direct rating (DR) task in which respondents state the value of each swing either by allocating a fixed number of points (usually 100 points) between the 'swings', or by directly rating each swing on a standard point scale with the top ranked swing automatically receiving the maximum possible number points (usually 100 points) [26, 27]. The relative importance weights of each ranked swing can then be calculated using the proportion of points given to each swing [28, 29].

Both DCEs and SW produce relative attribute ranks and weights and share a similar theoretical foundation [30]. However, there are notable theoretical and practical differences between the two techniques. First, SW does not comply with the 'random' aspect of RUT as choices in SW are assumed to be deterministic in nature. Second, SW directly captures attribute weights at an individual level while in a DCE the relative importance scores are estimated as a secondary outcome after the development of an econometric model. Third, SW does not

require a formal experimental design and can thus use smaller sample sizes than are required for DCE studies [28]. Finally, the cognitive demand of a DCE is believed to be higher than that of a SW task [28, 31].

While both DCE and SW have been implemented in healthcare preference research, empirical evidence directly comparing DCE and SW outcomes in terms of attribute relative importance and ease of comprehension and completion is lacking [31]. This study aimed to address this gap in knowledge by empirically comparing DCE and a SW-DR derived attribute relative importance rankings and weights.

METHODS

Study context and ethics

The outcomes of a study assessing the preferences of Non-Small Cell Lung Cancer (NSCLC) patients in Italy and Belgium for treatment was used for this comparative analysis. Details on the study design have been published elsewhere [32, 33]. This case study was identified as suitable for the comparison of DCE and SW-DR due to the potentially fragile physical state or diminished cognitive status of the patient [34-37]. The study was approved by the Ethical Committee of the European Institute of Oncology IRCCS (IEO, Milan, Italy; reference R1142/20-IEO 1206) and the "Ethische Commissie Onderzoek UZ/KU Leuven" (Belgium; reference S63007).

Respondents and recruitment

Patients with NSCLC were recruited through clinical partners in Italy and Belgium. Respondents were selected and referred to the PREFER research team by the treating oncologists at cancer treatment centers in Belgium and in Italy [33]. To be eligible patients had to understand Italian or Dutch, be 18 years or older, and have a histological or cytological diagnosis of NSCLC as evaluated by clinicians. Patients were not eligible if they (as evaluated by the clinician): i) had cognitive impairments rendering the participant incapable of informed consent or ii) were unable to understand the study materials.

Attribute and level selection

Attributes and levels were identified and refined according to best practices and guidelines [38-41]. This included a literature review, six nominal group-technique based focus groups in Italy and Belgium with NSCLC patients [42, 43], and a multi-stakeholder discussion with clinicians and preference experts [44]. Five attributes with three levels each were identified as relevant for the study (see Table 1).

DCE experimental design

A Bayesian D-efficient design consisting of two-unlabeled-alternative forced-choice tasks was constructed for the DCE using Ngene (ChoiceMetrics. Sydney, Australia)[19, 45]. A total of 36 unique choice tasks were generated which were divided over three 12-choice task blocks. Respondents were randomly assigned to complete one of those blocks. Attribute prior information for DCE design optimization was generated using previously published literature and best guesses. The survey was pilot tested among respondents in Italy (N=50) with the outcomes of a conditional logit model used to inform the final experimental design. Interactions between the attributes '5-year survival' and respectively 'Risk of long-lasting skin problems', 'risk of extreme tiredness' and 'mode of administration' were accounted for in this design. An example of a DCE choice task can be found in Figure 1.

Table 1. Attributes and levels included in the DCE and the swings used in the SW

Attributes	**Levels**
How the treatment is being given to you (Mode)	Oral treatment Intravenous infusion lasting 24 hours Intravenous infusion lasting 12 hours
Chance of surviving 5 years after beginning of the cancer treatment (5-year survival)	10% 20% 40%
Risk of persistent skin problems (Skin problems)	10% 20% 40%
Risk of being extremely tired (Tiredness)	10% 40% 60%
Severity of hair loss (Hair)	No hair loss Weakening/Thinning of the hair Complete loss of hair

Swings	**Worst**	→	**Best**
How the treatment is being given to you	Intravenous infusion lasting 24 hours	→	Oral treatment
Chance of surviving 5 years after beginning of the cancer treatment	10%	→	40%
Risk of persistent skin problems	40%	→	10%
Risk of being extremely tired	60%	→	10%
Severity of hair loss	Complete loss of hair	→	No hair loss

SW Design

A SW-DR task was developed using the attributes and levels identified for the DCE. In the SW section, respondents were asked to choose which attribute they preferred to swing from the lowest (worst) to the highest (best) level first. Respondents were asked to rank all other swings subsequently from most to least preferred. The order in which the swings were presented

was randomized in this section. In the DR section, respondents were asked to rate each of the swings relative to the others by giving it between 0 and 100 points with the exception of the highest ranked swing, which automatically received 100 points [26, 27]. Respondents were instructed on what this relative rating means as follows: 'if you give 50 points to improve a feature, it means that you think improving it is half as important as improving the top ranked attribute because you gave it half as many points.' This unrestricted valuation is assumed to be simpler for respondents than PA from a fixed pool and has been found to be more reliable than restricted PA methods [46-49], making it more suitable for this study population who may have more fragile physical states or diminished cognitive status [34-37]. An example SW-DR task can be found in Figure 1.

Survey

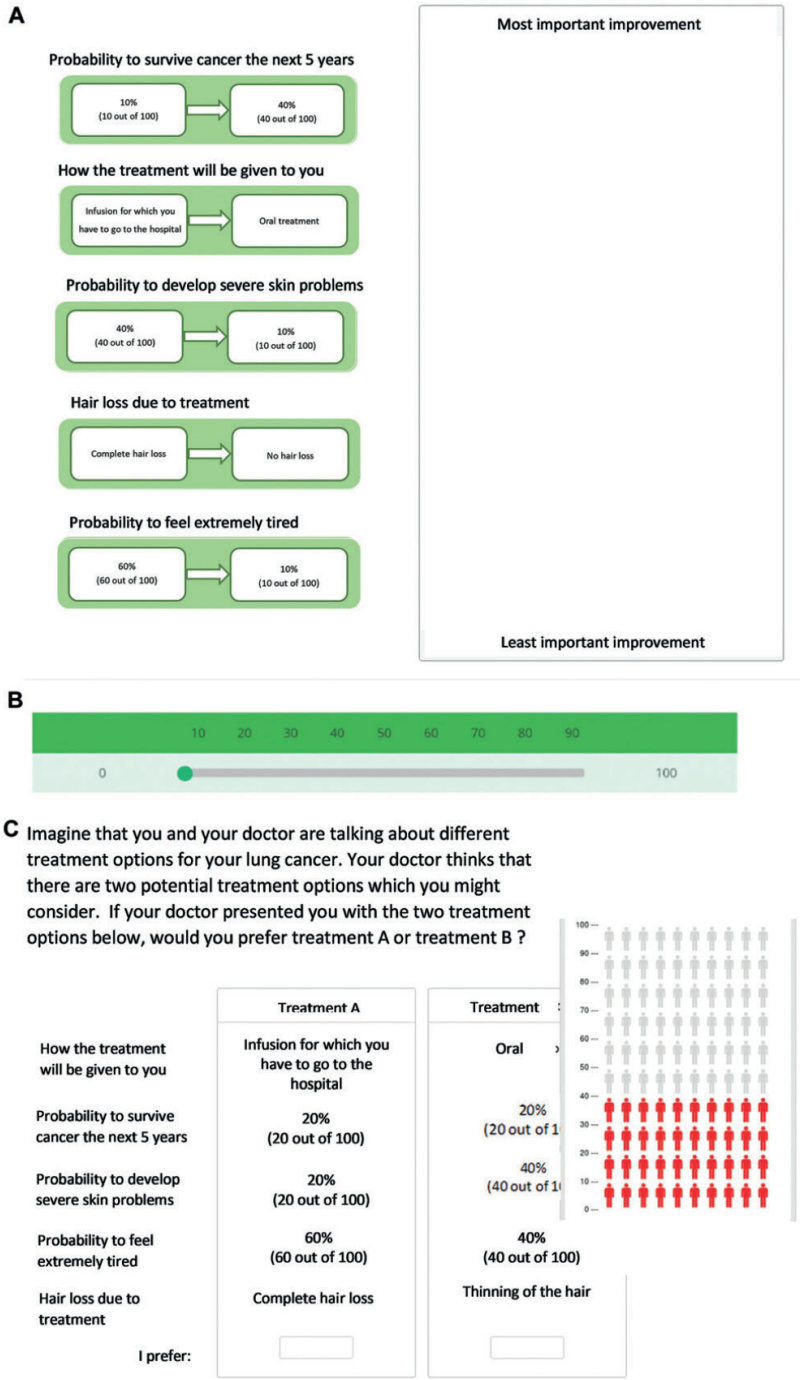
Both the DCE and SW-DR tasks were included as parts of a one-time online survey with respondents able to pause and return to the survey. The survey was programmed in Sawtooth software (lighthouse studio 9.13) and consisted of six parts. First, respondents were informed about the study and provided consent for data collection prior to answering socio-demographic and medical history related questions. Second, respondents watched two different educational videos consisting of text and animations with voiceovers giving (1) an introduction with information on Lung Cancer and a detailed descriptions of the attributes and levels included, and (2) instructions on how to complete the first choice task. Third, respondents were randomly assigned to receive either the DCE or SW task first to avoid any ordering effects. Fourth, respondents completed quality of life related questions (EQ-5D) [50, 51]. Fifth, respondents watched a video with instructions on how to complete the second choice task. Finally, respondents were asked to complete psychosocial measures including measures of health literacy [52, 53].

After each choice task, respondents were asked two feedback questions about ease of understanding and answering the choice tasks on a 5-point Likert scale ranging from very easy to very difficult. The survey was pre-tested with five LC patients in think-aloud interviews.

Statistical analysis

Only completed surveys were included in the analysis. One respondent was excluded from the data set due to flatlining behavior. Statistical analysis was performed with Nlogit version 6 and R version 4.0.4 [54]. A significance level of $p < 0.05$ was used for all analyses.

Figure 1. Illustration of survey elements: A. SW ranking task; B. SW direct ranking task (respondents were asked to put a relative weight for all swings in attributes); C. DCE choice task task (pop-up shown to explain risk attribute (pop-up shown to explain risk attribute



Respondent background characteristics and feedback questions

Respondent background characteristics (including general demographic and medical history information) were categorized and are presented as counts with percentages. Frequencies and chi-square tests were conducted to compare feedback of respondents regarding their perceived difficulty in understanding and answering the DCE and SW questions.

DCE analysis

Random Parameter Logit models (RPLs) were constructed to adjust for the multilevel structure of the data. [12, 16]. All risk and benefit attributes were assumed to be linear and categorical attributes were dummy coded. The significance level of the standard deviation of the attributes was used to test which attributes should be included in the final model as random parameters to account for preference heterogeneity. The systematic utility component (V), which describes the measurable utility of a specific treatment based on the attributes included in the DCE, was tested in both countries separately using the equation below. The β1- β7 coefficients in the equation represent the attribute level estimates indicating the relative importance of each attribute level for individual i.

$V = \beta_{1,i} *$ Mode infusion at hospital for 12 hours $+ \beta_{2,i} *$ Mode infusion at hospital for 24 hours $+ \beta_{3,i} *$ 5-year survival $+ \beta_{4,i} *$ risk of long-lasting skin problems $+ \beta_{5,i} *$ risk of extreme tiredness $+ \beta_{6,i}$ Hair loss some loss $+ \beta_{7,i}$ Hair loss no loss

Pre-specified interaction terms that significantly contributed to model fit (as assessed using a Log likelihood ratio test (LL ratio)) were included in the model. A 'choice task-order' variable was included in the model as an interaction term with the attribute levels to test whether the task order (i.e., DCE first or SW first) influenced the outcomes. All analyses were performed separately for data from Italy and Belgium. Individual specific conditional parameter estimates were estimated for each respondent using the final model. Individual attribute weights and rankings were calculated with these parameter estimates and averaged to estimate the mean population weights and rankings.

SW analysis

The SW analysis was performed by analyzing the patients' rankings of the attributes and the points allocated to the different attributes. The individual attribute relative importance weights were calculated using both the rank ordered centroid (ROC) weight method and the DR weight method per patient. The ROC weight method calculates a relative weight representing the distance between adjacent ranks on an ordinal or normalized scale [55].

The ROC weight for an attribute with rank i equals (in case of 5 attributes):

$$w_i = \frac{1}{5}\sum_{n=i}^{5}\frac{1}{n}, \qquad i = 1, \ldots, 5.$$

The DR method is used to generate individual proportional weights for an attribute with rank i and allocated points pi and equals (in case of 5 attributes):

$$w_i = \frac{p_i}{\sum_{i=1}^{5} p_i}, \qquad i = 1, \ldots, 5.$$

These individual weights were averaged over all patients per country to obtain the average weights which are the equivalent of the attribute relative importance scores resulting from DCEs.

Comparison of relative importance ranking and weights between methods

The rankings and weights resulting from the DCE, SW-ROC and SW-DR were quantitatively compared. Ranking agreement was evaluated with Cohens's weighted kappa which measures inter-rater reliability while accounting for chance similarities in rating [56, 57]. Differences in the ranking between the DCE and SW-ROC were analyzed and tested with an ordered logit model [58]. Dirichlet regression was used to analyze whether the relative attribute importance weights differed between methods [59]. Dirichlet regression models can be regarded as a generalization of beta regression models for proportions and percentages and are particularly suited for the analysis of compositional data (i.e., for weights that add up to 1) [60]. In a Dirichlet regression model the attribute weights are assumed to be distributed with a Dirichlet distribution with parameters μi, i = 1,…, 5, mean attribute weights that add up to one, and a precision parameter φ (according to the so-called alternative parametrization [61]). The mean attribute weights are modelled with a logit link function similar to logistic regression:

$$logit(\mu_i) = \eta_i = \beta_{0,i} + \beta_{1,i} D_{SW}, \; i = 1,\ldots,5$$

Here the logit of μ for individual i is equal to the linear predictor η and is modelled with an intercept β_(0,i), representing the DCE and with a dummy variable DSW for the method as covariate. We defined the attribute 5-year survival as the base category, with β_(0,survival)= β_(1,survival)=0. In this way the corresponding values of μi equal:

$$\mu_i = \frac{e^{\eta_i}}{\sum_{j=1}^{5} e^{\eta_i}} \text{ and } \mu_{survival} = \frac{1}{\sum_{j=1}^{5} e^{\eta_i}}.$$

The precision parameter is modelled with a log link function with method as covariate:

$$log(\varphi) = \alpha_0 + \alpha_1 D_{SW}.$$

The parameter estimates β1,i can be interpreted as odds ratios after exponentiation relative to survival as base category [60]. Maximum likelihood estimation is used for obtaining the parameter estimates [62]. Finally, covariates were added to the models to correct for possible effects of method, for educational level, health literacy, gender, age, cancer stage and treatment history.

Table 2. Demographic characteristics of the sample

		ITALY (N=158)		BELGIUM (N=149)	
		N	%	N	%
Sex	Male	88	55.7	89	59.7
	Female	70	44.3	60	40.3
Age at survey completion	< 71	102	64.6	104	69.8
	≥ 71	56	35.4	45	30.2
Education	No degree	0	0	6	4
	Primary school	12	7.6	6	4
	Middle school	37	23.4	30	20.1
	Secondary school	51	32.3	51	34.2
	Professional degree	19	12	26	17.4
	Bachelor's degree	4	2.5	0	0
	Master's degree	26	16.5	14	9.4
	Postgraduate degree	5	3.2	2	1.3
	Other	4	2.5	14	9.4
Family & relationship status	Single no kids	15	9.5	18	12.1
	Single with kids	12	7.6	17	11.4
	Partner with kids	64	40.5	38	25.5
	Partner no kids	67	42.4	76	51
Family history	Yes	45	28.5	37	24.8
	No	99	62.7	97	65.1
	Don't know	14	8.9	15	10.1
Cancer stage	I, II	78	49.4	65	43.6
	III, IV	80	50.6	84	56.4
Type of treatment	No treatments	21	13.3	0	0
	Surgery	94	59.5	78	52.3
	Chemotherapy	55	34.8	88	59.1
	Immunotherapy	35	22.2	78	52.3
	Radiotherapy	35	22.2	46	30.9
	Other	18	11.4	12	8.1
	Don't know	3	1.9	0	0
Lines of treatment	No treatment	72	45.6	72	48.3
	1 treatment	34	21.5	14	9.4
	2 treatments	14	8.9	15	10.1
	3 treatments	17	10.8	48	32.2
	> 3 treatments	21	13.3	0	0
Age when diagnosed	<55	28	17.7	21	14.1
	55-64	48	30.4	57	38.2
	65-74	57	36.1	59	39.6
	>=75	25	15.8	12	8.1
Health literacy (Newest Vital Sign)	Very Limited literacy	7	4.4	12	8.1
	Limited literacy	40	25.3	32	21.5
	Adequate literacy	111	70.3	105	70.5

RESULTS

Demographics

A sample of N=307 NSCLC patients was obtained from an N=560 requests to patients (N=159 declined invite; N=94 withdrew consent). No significant differences were found between the countries in respondents' gender, age, cancer stage, or family history of cancer. Respondents in both countries significantly differed in family and relationship status ($\chi 2$ (3) = 8.1, p = .045), education level ($\chi 2$ (2) = 7.248, p = .027), and health literacy (t (305) = -6.591, p < .001). Patient demographic information can be found in Table 2.

Choice Task Feedback

Most respondents found the DCE and SW tasks very easy or easy to understand and answer (74.6% and 64.5% for DCE and 73.0% and 69.7% for SW respectively), with no differences between country found. The ease of understanding and answering the DCE and understanding the SW task was related to educational level with those higher levels of education reporting greater ease (P < ****). No association was found for health literacy.

Preferences based on DCE

Appendix A Table 1 shows the outcomes of the DCE for respondents from Italy and Belgium. Respondents in both countries preferred treatments with a higher probability of 5-year survival, lower Risk of long-lasting skin problems and a lower Risk of extreme tiredness. Patients in both countries preferred some hair loss and no hair loss over complete loss of hair. Only Italian respondents significantly preferred oral treatment over infusions. Significant heterogeneity was found for all attributes in both countries.

Attribute ranking based on SW

Appendix A Table 2 shows the outcomes the SW-ROC and SW-DR analysis for Italy and Belgium. For respondents in both countries 5-year survival was the most important attribute (ROC weight = 0.43 and 0.42 for Italy and Belgium respectively), followed by Risk of extreme tiredness (ROC weight = 0.18 and 0.20 for Italy and Belgium respectively). The least important attribute was hair loss (ROC weight = 0.10 and 0.11 for Italy and Belgium respectively). Heterogeneity in preferences for the different attributes was evidenced in the standard deviations of the individual weights which ranged between 0.06 and 0.12.

Comparing relative importance of the attribute between DCE and SW-DR

Table 3 shows ranks and relative importance weights for the two countries and two methods separately. For 89.3% of the Belgian respondents 5-year survival was the most important

attribute based on the DCE compared to 85.2% for SW. Similarly, for 94.5% of the Italian respondents 5-year survival was the most important attribute based on the DCE compared to 90.5% for SW. Although the attribute 5-year survival was weighted highest and therefore most important across both methods and countries, the weights of all the attributes and their relative importance differed substantially between the two methods in both countries. The largest difference was found for the weight of '5-year survival' which was much greater in the DCEs (59-63% of total weight) than in the SW methods (31-33%). The differences in the average weights ranking order are evidenced in their 95% confidence intervals which minimally overlap between methods (see Table 3). The less important had different weights, but were more comparable across methods.

The agreement between the ranking based on ROC and DR showed high agreement (ρ = 0.91, CI (0.89-0.93) for Italy and ρ = 0.90 (0.88-0.93) for Belgium). Agreement between the ranking of the DCE and SW-ROC and SW-DR was moderate with weighted Kappa correlation coefficients varying between 0.53 and 0.55. Despite the similar preference weights for the 5-year survival and Tiredness attributes, the overall ranking of the attributes differed significantly between DCE and SW – ROC tasks for both countries ($\chi 2$ = 2042.9, 4 df, $p < .0001$ for Italy; $\chi 2$ = 1932.5, 4 df, $p < .0001$ for Belgium). For the Italian respondents the attributes of Mode and Hair swapped their rank (3rd or 5th) based on the method from which the ranking was derived. For the Belgian respondents, the attributes of Mode, Skin problems, and Hair changed ranking between being 3rd, 4th, or 5th most important attribute.

According to the Dirichlet regression models (see Table 4, Appendix C) the differences between DCE and SW were significant (LL ratio = 466.4 for Italy, $p < .0001$; LL ratio = 435.0 for Belgium, $p < .0001$). Weights of the SW were more equally apportioned over the included attributes as compared to the DCE, where the majority of the weight was allocated to the 5-year survival attribute (Figure 2). Relative to survival, the weights based on the SW for skin problems, mode of administration, tiredness and hair problems were significantly larger compared to the DCE weights ($p<0.001$; see Figure 2). Moreover, for Italy the weights based on the SW were significantly less dispersed compared to the DCE (φ = 0.75, CI: 0.64 – 0.87; $p < 0.001$). These differences remained highly significant even after correcting for educational level, health literacy, gender, age, cancer stage and treatment experience.

Table 3. Rank and Attribute weights (95% confidence interval) representing their relative importance based on DCE and SW-DR separately for Italy and Belgium.

	Italy				Belgium			
	DCE		SW-DR		DCE		SW-DR	
	Rank	Weight	Rank	Weight	Rank	Weight	Rank	Weight
Mode of administration	5	0.05 (0.04- 0.06)	3	0.18 (0.15 – 0.19)	5	0.02 (0.02 – 0.02)	4	0.16 (0.14 – 0.17)
5-Year Survival	1	0.63 (0.61 – 0.66)	1	0.33 (0.31 – 0.34)	1	0.59 (0.57 – 0.62)	1	0.31 (0.30 – 0.33)
Risk of long-lasting skin problems	4	0.08 (0.07 – 0.08)	4	0.16 (0.15 – 0.17)	4	0.08 (0.08 – 0.09)	3	0.18 (0.17 – 0.19)
Risk of extreme tiredness	2	0.16 (0.14 – 0.17)	2	0.19 (0.18 – 0.20)	2	0.20 (0.18 – 0.22)	2	0.22 (0.21 – 0.23)
Hair loss	3	0.08 (0.08 – 0.09)	5	0.14 (0.13 – 0.15)	3	0.10 (0.09 – 0.11)	5	0.13 (0.12 – 0.14)

Table 4. Dirichlet regression beta-parameters ORs (SW compared to DCE and 95% CI) and the dispersion parameter (Ln Phi) showing SW derived weights are significantly higher than DCE derived weights.

		5- year survival	Mode of administration	Risk of long-lasting skin problems	Risk of extreme tiredness	Hair loss	Ln Phi
Italy	OR	ref	4.87	3.80	2.46	2.58	0.75
	(95% CI)		(4.13, 5.74)	(3.24, 4.45)	(2.14, 2.83)	(2.215, 3.03)	(0.64, 0.87)
Belgium	OR	ref	6.41	3.41	2.15	1.67	1.02
	(95% CI)		(5.27, 7.80)	(2.90, 4.01)	(1.87, 2.48)	(1.42, 1.98)	(0.87, 1.20)

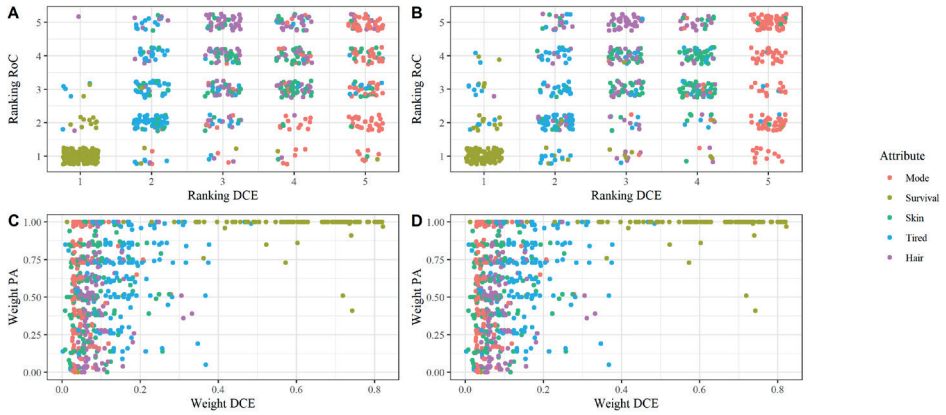


Figure 2. Comparison of the rankings derived from the ROC and DCE in Italy (A) and Belgium (B), and the attribute weighting from the point task and the DCE in Italy (C) and Belgium (D)

DISCUSSION

This study used empirical evidence to compare the relative importance of NSCLC treatment profile attribute ranking and weighting when assessed using a DCE or SW task in a study of Italian and Belgian patients. In general, respondents indicated that both DCE and SW tasks were easy or very easy to understand and answer, supporting their use in future research on treatment preferences in similar patient populations. However, significant differences were found in the relative importance and the ranking of the attributes between the SW and the DCE. Moreover, which aspect of the SW method was used (ROC or DR) appeared to influence both ranking and weight of the attributes.

The difference in relative weights and ranking may be in part due to the differences in how the two methods assess patient preferences and how respondents engage with the tasks. In an optimally designed DCE, respondents are forced to make trade-offs between attributes when choosing an alternative in each choice task and cannot directly state their individual attribute valuations. This is due to the multi-attribute nature of a DCE where the total utility of all attributes should guide decision-making which better reflects real-world decision making and may result in more realistic point estimates for relative valuations [18]. In the DR method used in our SW task, there is no trade-off as the method is unrestricted allowing respondents to assign any number of points to attributes (excluding the most important attribute which automatically receives 100 points) [26]. This may result in relative importance weights which are more equivalent leading to a relative undervaluing of the more important attributes and overvaluing of the less important attributes [48]. The reduced spread of the importance scores may paint a more homogeneous picture of attribute relative importance than is the case.

Additionally, since attribute ranks and weights valuations in a DCE are derived from an econometric model rather than a direct assessment of value (as is the case in a SW-DR task), they are more sensitive to modelling decisions researchers make when deriving the attribute preference estimates [31, 63].

To explore if a restricted PA task (which more closely resembles the tradeoffs in a DCE) results in more equivalent relative importance scores than the unrestricted DR task, a small post-hoc add-on study was conducted (see Appendix B). In this study 14 (randomly selected) Italian patients who previously completed the full survey were asked to complete the SW-DR task from the original survey as well as an additional restricted PA task. Respondents were asked to divide a total of 100 points over five attributes rather than simply rate each swing on a 100-point scale thus forcing respondents to trade-off when allocating points to the five attributes [28]. While small and underpowered due to the explorative nature of this study, the results indicate that weights based on this restricted PA task more closely resemble the DCE study outcomes than those from the unrestricted DR task which replicate previous findings [48] (see Appendix B). Further studies are needed to confirm if findings from this exploratory analysis hold with larger samples, different sample composition and different choice contexts to see whether the differences remain and how comparable the outcomes are to DCE outcomes.

A primary strength of this study is that the empirical evidence used to compare the two methods was generated in a one-time survey of NSCLC patients who completed both methods allowing for direct comparison of results. The within-subjects design reduced the chance of confounding factors playing a role in different preference outcomes. This survey was developed after an extensive qualitative study in close collaboration with a multi-disciplinary team of clinicians, patients, and researchers. The tasks were explained using informational videos designed for the study and the online setting allowing respondents to pause the educational material or the survey and return to it at a later time in. The online setting also allowed for multi-country, location independent data collection and access for those with more serious disease complications or fatigue to participate, increasing the generalizability of the findings to other NCSLC populations and reducing the chance of bias.

However, this study also had some limitations. First, SW tasks were originally designed to be conducted in-person via a trained facilitator [28, 64]. The current study was administered online with respondents completing the survey on their own. While online surveys are less costly and time-consuming than interviewer-led studies and SW surveys have previously been done online, the presence of an interviewer allows for assistance and clarification of questions or issues which could arise while the participant is completing the choice task [65]. This can be especially helpful when attributes are complex or the target population experiences cognitive impairments [64]. The patient feedback questions indicated that the online setting was not a problem for this study. Second, the sample was composed of relatively old and 'fragile' NSCLC patients, reducing generalizability to younger or less 'fragile' patient populations.

Generalizability is also limited by the fact that the digital format of the survey may have discouraged those patients with lower digital literacy from participating as well as those who lack access to computer equipment or to the internet [66]. Finally, it is unclear whether patients received support from relatives while completing the survey. If this occurred, those supporting the patient in completing the survey could have influenced the outcomes of the survey such that the values measured did not solely reflect the true values of the patient.

In conclusion, this study found significant differences in attribute importance between DCE and SW-DR as well as a greater spread in the DCE-derived relative importance of the attributes. The latter may indicate that DCEs are more suitable to producing attribute importance information as they force respondents to trade-off between attributes. This yields results that more likely reflect the real-world valuations that patients have for these attributes. However, further studies confirming these findings as well as SW studies with restricted PA tasks are warranted to provide accurate guidance for methods selection when studying relative attribute importance across a wide array of preference-relevant decisions.

REFERENCES

1. de Bekker-Grob, E.W., et al., Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public-Private Project. Patient, 2017. 10(3): p. 263-266.
2. Ho, M., et al., A Framework for Incorporating Patient Preferences Regarding Benefits and Risks into Regulatory Assessment of Medical Technologies. Value in Health, 2016. 19(6): p. 746-750.
3. Marsh, K., et al., Health Preference Research in Europe: A Review of Its Use in Marketing Authorization, Reimbursement, and Pricing Decisions-Report of the ISPOR Stated Preference Research Special Interest Group. Value Health, 2020. 23(7): p. 831-841.
4. Ryan, M., K. Gerard, and M. Amaya-Amaya, Using discrete choice experiments to value health and health care. Vol. 11. 2007: Springer Science & Business Media.
5. Marsh, K., et al., How to integrate evidence from patient preference studies into health technology assessment: a critical review and recommendations. International Journal of Technology Assessment in Health Care, 2021. 37(1).
6. Soekhai, V., et al., Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discovery Today, 2019. 24(7): p. 1324-1331.
7. Smith, I.P., et al., Methodological Priorities for Patient Preferences Research: Stakeholder Input to the PREFER Public–Private Project. The Patient - Patient-Centered Outcomes Research, 2021.
8. Soekhai, V., et al., Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discov Today, 2019. 24(7): p. 1324-1331.
9. Lancsar, E. and J. Louviere, Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics, 2008. 26(8): p. 661-77.
10. Ryan, M., K. Gerard, and M. Amaya-Amaya, Using Discrete Choice Experiments to Value Health and Health Care. The Economics of Non-Market Goods and Resources, ed. I.J. Bateman. 2008, Dordrecht: Springer.
11. de Bekker-Grob, E.W., M. Ryan, and K. Gerard, Discrete choice experiments in health economics: a review of the literature. Health Econ., 2012. 21(2): p. 145-72. doi: 10.1002/hec.1697. Epub 2010 Dec 19.
12. Hensher, D., J.M. Rose, and W.H. Greene, Applied Choice Analysis: second edition. 2015, Cambridge, UK: Cambridge University Press.
13. Louviere, J.J., D.A. Hensher, and J.D. Swait, Stated Choice Methods; analysis and application. 2000, Cambridge: Cambridge University Press.
14. McFadden, D., The Choice Theory Approach to Market Research. Marketing Science, 1986. 5(4): p. 275–297.
15. Viney, R., E. Lancsar, and J. Louviere, Discrete choice experiments to measure consumer preferences for health and healthcare. Expert Review of Pharmacoeconomics & Outcomes Research, 2002. 2(4): p. 319-326.
16. Hauber, A.B., et al., Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value Health, 2016. 19(4): p. 300-15.
17. Gonzalez, J.M., A Guide to Measuring and Interpreting Attribute Importance. The Patient - Patient-Centered Outcomes Research, 2019. 12(3): p. 287-295.
18. De Bekker-Grob, E.W., et al., Are Healthcare Choices Predictable? The Impact of Discrete Choice Experiment Designs and Models. Value in Health, 2019. 22(9): p. 1050-1062.

19. Rose, J.M. and M.J.C. Bliemer, Constructing efficient stated choice experimental desings. Transport Reviews, 2009. 9: p. 1-31.

20. Watson, V., F. Becker, and E. de Bekker-Grob, Discrete Choice Experiment Response Rates: A Meta-analysis. Health Econ, 2017. 26(6): p. 810-817.

21. DeShazo, J.R. and G. Fermob, Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. Journal of Environmental Economics and Management, 2011. 44(1): p. 123-143.

22. de Bekker-Grob, E.W., et al., Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. Patient, 2015. 8(5): p. 373-84.

23. Street, D.J., et al., Designing Discrete Choice Experiments for Health Care. 2008, Springer Netherlands. p. 47-72.

24. Whichello, C., et al., Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison. BMC Med Inform Decis Mak, 2020. 20(1): p. 114.

25. Friend, A.J., G.A. Ayoko, and H. Guo, Multi-criteria ranking and receptor modelling of airborne fine particles at three sites in the Pearl River Delta region of China. Sci Total Environ, 2011. 409(4): p. 719-37.

26. Mustajoki, J., R.P. Hamalainen, and A. Salo, Decision Support by Interval SMART/SWING-Incorporating Imprecision in the SMART and SWING Methods. Decision Sciences, 2005. 36(2): p. 317-339.

27. Jyri, M., R.P. Hämäläinen, and A. Salo, Decision support by interval SMART/SWING—incorporating imprecision in the SMART and SWING methods. Decision Sciences 2005. 35(2): p. 317-339.

28. Marsh, K., et al., Multiple Criteria Decision Analysis for Health Care Decision Making--Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health, 2016. 19(2): p. 125-37.

29. Németh, B., et al., Comparison of weighting methods used in multicriteria decision analysis frameworks in healthcare with focus on low- and middle-income countries. Journal of Comparative Effectiveness Research, 2019. 8(4): p. 195-204.

30. von Neumann, J. and O. Morgenstern, Theory of Games of Economic Behavior. 2nd ed. 1944, Princeton, NJ: Princeton University Press.

31. Tervonen, T., et al., MCDA swing weighting and discrete choice experiments for elicitation of patient benefit-risk preferences: a critical assessment. Pharmacoepidemiology and Drug Safety, 2017. 26(12): p. 1483-1491.

32. Durosini, I., et al., Patient Preferences for Lung Cancer Treatment: A Qualitative Study Protocol Among Advanced Lung Cancer Patients. Front Public Health, 2021. 9: p. 622154.

33. Monzani, D., et al., Patient Preferences for Lung Cancer Treatments: A Study Protocol for a Preference Survey Using Discrete Choice Experiment and Swing Weighting. Frontiers in Medicine, 2021. 8(1107).

34. Simó, M., et al., Cognitive and Brain Structural Changes in a Lung Cancer Population. Journal of Thoracic Oncology, 2015. 10(1): p. 38-45.

35. Eggen, A.C., et al., Factors associated with cognitive impairment and cognitive concerns in patients with metastatic non-small cell lung cancer. Neuro-Oncology Practice, 2021.

36. Grosshans, D.R., et al., Neurocognitive function in patients with small cell lung cancer. Cancer, 2008. 112(3): p. 589-595.

37. Whitney, K.A., et al., Is"chemobrain" a transient state? A prospective pilot study among persons with non-small cell lung cancer. J Support Oncol, 2008. 6(7): p. 313-21.

38. Bridges, J.F.P., et al., Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health, 2011. 14(4): p. 403-413.

39. Hensher, D.A., et al., Applied choice analysis: a primer. 2005: Cambridge university press.

40. Hollin, I.L., et al., Reporting Formative Qualitative Research to Support the Development of Quantitative Preference Study Protocols and Corresponding Survey Instruments: Guidelines for Authors and Reviewers. The Patient - Patient-Centered Outcomes Research, 2020. 13(1): p. 121-136.

41. Coast, J. and S. Horrocks, Developing attributes and levels for discrete choice experiments using qualitative methods. Journal of health services research & policy, 2007. 12(1): p. 25-30.

42. Hiligsmann, M., et al., Nominal group technique to select attributes for discrete choice experiments: an example for drug treatment choice in osteoporosis. Patient Prefer Adherence, 2013. 7:133-9.

43. Janssens, R., et al., Key Determinants of Health-Related Quality of Life Among Advanced Lung Cancer Patients: A Qualitative Study in Belgium and Italy. Frontiers in Pharmacology, 2021. 12(2464).

44. Reed Johnson, F., et al., Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value in Health, 2013. 16(1): p. 3-13.

45. Bliemer, M.C.J. and J.M. Rose, Efficiency and sample size requirements for stated choice experiments, in Transportaion Research Broad Annual Meeting. 2009: Washington DC.

46. Bottomley, P.A., J.R. Doyle, and R.H. Green, Testing the Reliability of Weight Elicitation Methods: Direct Rating versus Point Allocation. Journal of Marketing Research, 2000. 37(4): p. 508-513.

47. Van Til, J.A., et al., The Use of Multi-Criteria Decision Analysis Weight Elicitation Techniques in Patients with Mild Cognitive Impairment. The Patient: Patient-Centered Outcomes Research, 2008. 1(2): p. 127-135.

48. Doyle, J.R., R.H. Green, and P.A. Bottomley, Judging Relative Importance: Direct Rating and Point Allocation Are Not Equivalent. Organizational Behavior and Human Decision Processes, 1997. 70(1): p. 65-72.

49. Bottomley, P.A. and J.R. Doyle, A comparison of three weight elicitation methods: good, better, and best. Omega, 2001. 29(6): p. 553-560.

50. Brooks, R., EuroQol: the current state of play. Health Policy, 1996. 37(1): p. 53-72.

51. Koide, R., et al., Quality assessment using EQ-5D-5L after lung surgery for non-small cell lung cancer (NSCLC) patients. General Thoracic and Cardiovascular Surgery, 2019. 67(12): p. 1056-1061.

52. Chew, L.D., K.A. Bradley, and E.J. Boyko, Brief questions to identify patients with inadequate health literacy. health, 2004. 11: p. 12.

53. Powers, B.J., J.V. Trinh, and H.B. Bosworth, Can This Patient Read and Understand Written Health Information? JAMA, 2010. 304(1): p. 76.

54. "R Core Team, R.F.f.S.C., R: A Language and Environment for Statistical Computing. 2017: Vienna, Austria.

55. Danielson, M. and L. Ekenberg. Trade-offs for ordinal ranking methods in multi-criteria decisions. in International Conference on Group Decision and Negotiation. 2016. Springer.

56. Cohen, J., Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 1968. 70(4): p. 213.

57. Landis, J.R. and G.G. Koch, The Measurement of Observer Agreement for Categorical Data. Biometrics, 1977. 33(1): p. 159.

58. Fok, D., R. Paap, and B. Van Dijk, A RANK-ORDERED LOGIT MODEL WITH UNOBSERVED HETEROGENEITY IN RANKING CAPABILITIES. Journal of Applied Econometrics, 2012. 27(5): p. 831-846.

59. Gueorguieva, R., R. Rosenheck, and D. Zelterman, Dirichlet component regression and its applications to psychiatric data. Computational Statistics & Data Analysis, 2008. 52(12): p. 5344-5355.

60. Ferrari, S. and F. Cribari-Neto, Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics, 2004. 31(7): p. 799-815.

61. Douma, J.C. and J.T. Weedon, Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods in Ecology and Evolution, 2019. 10(9): p. 1412-1430.

62. Maier, M., DirichletReg: Dirichlet Regression for Compositional Data in R. Research Report Series / Department of Statistics and Mathematics, 2014. 125.

63. Rakotonarivo, O.S., M. Schaafsma, and N. Hockley, A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. J Environ Manage, 2016. 183: p. 98-109.

64. Tervonen, T., et al., MCDA swing weighting and discrete choice experiments for elicitation of patient benefit-risk preferences: a critical assessment. Pharmacoepidemiol Drug Saf, 2017. 26(12): p. 1483-1491.

65. Aubert, A.H., F. Esculier, and J. Lienert, Recommendations for online elicitation of swing weights from citizens in environmental decision-making. Operations Research Perspectives, 2020. 7: p. 100156.

66. Oliveri, S., et al., Opportunities and Challenges of Web-Based and Remotely Administered Surveys for Patient Preference Studies in a Vulnerable Population. Patient Preference and Adherence, 2021. Volume 15: p. 2509-2517.

APPENDIX

Appendix A Model outputs of the DCE and SW analyses

Table 1. Preference of Italian and Belgian NSCLC patients for treatment profiles, based on DCE

Italian data		**Mean**			**SD**		
		Estimate	*SE*	*95% CI*	*Estimate*	*SE*	*95% CI*
Mode of administration	Oral treatment (ref)	0			0		
	Infusion in hospital 12 hours	-0.65***	0.16	-0.97;-0.33	0.00	0.37	-0.75;0.76
	Infusion in hospital 24 hours	-0.64***	0.17	-0.97;0.31	0.87***	0.20	0.47;1.27
5-Year Survival		0.43***	0.05	0.33;0.54	0.22***	0.03	0.16;0.28
Risk of long-lasting skin problems		-0.04***	0.01	-0.06;-0.03	0.04***	0.01	0.02;0.05
Risk of extreme tiredness		-0.06***	0.01	-0.07;-0.04	0.04***	0.01	0.03;0.04
Hair loss	Complete loss of hair (ref)	0			0		
	Some hair loss	0.90***	0.16	0.58;1.21	0.20	0.30	-0.40;0.79
	No hair loss	1.41***	0.20	1.03;1.79	0.94***	0.24	0.48;1.40
Model fit measures:							
Log Likelihood		-697.40					
AIC		1422.8					
Pseudo R^2		0.47					
Belgian data		**Mean**			**SD**		
		Estimate	*SE*	*95% CI*	*Estimate*	*SE*	*95% CI*
Mode of administration	Oral treatment (ref)	0			0		
	Infusion in hospital 12 hours	-0.09	0.18	-0.43;0.25	0.45	0.33	-0.20;1.09
	Infusion in hospital 24 hours	-0.20	0.17	-0.54;0.14	0.55***	0.21	0.14;0.96
5-Year Survival		0.39***	0.05	0.30;0.47	0.23***	0.04	0.15;0.31
Risk of long-lasting skin problems		-0.04***	0.07	-0.06;-0.03	0.04***	0.01	0.02;0.05
Risk of extreme tiredness		-0.07***	0.01	-0.08;-0.05	0.05***	0.01	0.04;0.07
Hair loss	Complete loss of hair (ref)	0			0		

(cont'd)

Italian data		**Mean**			**SD**		
		Estimate	*SE*	*95% CI*	*Estimate*	*SE*	*95% CI*
	Some hair loss	1.16***	0.18	0.80;1.52	0.24	0.32	-0.39;0.86
	No hair loss	1.58***	0.24	1.10;2.05	1.12***	0.23	0.67;1.58
Model fit measures:							
Log Likelihood		-698.03					
AIC		1424.1					
Pseudo R^2		0.44					

***p<0.001

Table 2. Attribute ranks and weights (SD) representing their relative importance based on Swing Weighting (SW) (both Direct Rating (DR) and Rank Ordered Centroid (RoC)) by country (Italy and Belgium).

	Italy		**Belgium**	
	RoC	**DR**	**RoC**	**DR**
Mode of administration	3 0.16 (0.12)	3 0.18 (0.10)	3 0.14 (0.12)	4 0.16 (0.10)
5-Year Survival	1 0.43 (0.08)	1 0.33 (0.10)	1 0.42 (0.10)	1 0.31 (0.08)
Risk of long-lasting skin problems	4 0.14 (0.07)	4 0.16 (0.06)	4 0.14 (0.06)	3 0.18 (0.06)
Risk of extreme tiredness	2 0.18 (0.09)	2 0.19 (0.07)	2 0.20 (0.08)	2 0.22 (0.07)
Hair loss	5 0.10 (0.08)	5 0.14 (0.07)	5 0.11 (0.10)	5 0.13 (0.08)

Appendix B. Description and results of add-on study

A random sample of 14 respondents who initially participated in the study in Italy were recontacted. Upon providing informed consent, respondents were asked to read the instruction materials (including the educational videos) for the SW again. Respondents completed the SW and DR tasks as also included in the original survey. After that, new SW instructions were incorporated explaining the revised ranking exercise (point allocation (PA). Respondents were asked to complete this PA task by dividing 100 points over the five attributes reflecting their importance. Data of the SW-DR and SW-PA were analyzed in according to the analyses described in the main paper. Outcomes of the SW-DR and SW-PA were compared against each other as well as compared against the DCE outcomes of the original study.

Table 1 shows the outcomes of the SW-DR, the SW-PA and the DCE outcomes. On average we can see that the SW-PA weights for the attributes differ from those of the SW-DR and are more in resemblance with the weights obtained from the DCE.

Table 1. Relative attributes weights from the SW-DR, SW-PA of add-on study and DCE of original study

	SW-DR	SW-PA	DCE
Mode of administration	0.12	0.06	0.04
5-Year Survival	0.37	0.77	0.65
Risk of long-lasting skin problems	0.16	0.06	0.08
Risk of extreme tiredness	0.22	0.07	0.16
Hair loss	0.13	0.04	0.07

Appendix C. Ranked ordered logit model beta-parameters (SW rank compared to DCE) showing SW derived weights are significantly different than DCE derived weights iin some cases.Ŧ

		Hair loss	Mode of administration		Risk of long-lasting skin problems		5-year survival		Risk of extreme tiredness	
Italy	Mean,	Ref	2.17	***	1.12	***	0.09		-0.14	
	(S.E.)		(0.23)		(0.21)		(0.37)		(0.22)	
Likelihood ratio test X2 = 2042.9 (df = 4), p < 2.22e-16										
Belgium	Mean	ref	3.58	***	1.48	***	0.62	*	0.70	**
	(S.E.)		(0.35)		(0.22)		(0.31)		(0.23)	
Likelihood ratio test X2 = 1932.5 (df = 4), p < 2.22e-16										

Ŧ Note: the coefficients represent the effect of the method used to elicit the rank on the relative importance of the treatment attributes relative to Hair loss, a positive coefficient represents increasing importance while a negative coefficient represents decreasing importance. S.E.= Standard Error; *** p<0.0001, ** p <0.001, * p< 0.01

Chapter 8

The Impact of Video-Based Educational Materials with Voiceovers on Preferences for Glucose Monitoring Technology in Patients with Diabetes: A Randomized Study

Authors: Ian P Smith, Chiara L Whichello, Esther W de Bekker-Grob, Maureen P.M.H. Rutten-van Mölken, Jorien Veldwijk, G. Ardine de Wit

Status: Published

Smith, I.P., Whichello, C.L., de Bekker-Grob, E.W., Mölken, M.P.R.V., Veldwijk, J. and de Wit, G.A., 2023. The Impact of Video-Based Educational Materials with Voiceovers on Preferences for Glucose Monitoring Technology in Patients with Diabetes: A Randomised Study. The Patient-Patient-Centered Outcomes Research, pp.1-15.

ABSTRACT

Introduction: Ensuring patients have enough information about healthcare choices prior to completing a preference study is necessary to support the validity of the findings. Patients are commonly informed using text-based information with supporting graphics. Video-based information may be more engaging for the general patient population. This study aimed to assess 1) the impact that educating patients using video-based educational materials with a voiceover has on patient preferences compared to traditional text, and 2) whether this impact is consistent between two countries.

Materials and Methods: A video-based educational tool was developed to inform patients prior to completing a discrete choice experiment assessing preferences for glucose monitors. Patients with diabetes from the Netherlands and Poland were recruited through an online research panel. Respondents were randomized to receive information in either a text or a video with animations and a voiceover. Data was analyzed using a mixed-logit model.

Results: N=981 completed surveys were analyzed from the Netherlands (n=459) and Poland (n=522).

Differences were found between the countries, but no interpretable pattern of differences was found between the two types of educational materials. Patients spent less time in the educational material than would be necessary to fully review all of the content.

Conclusions: Simply providing educational material in a video with animations and voiceovers does not necessarily lead to better engagement from respondents or different preference outcomes in a sample of diabetes patients when compared to text. Increasing engagement with educational materials should be a topic of future research for those conducting patient preference research as no amount of educational material will be helpful if respondents do not access it.

1. INTRODUCTION

Patient preference information (PPI) is an increasingly used source of information to help develop and assess medical products before, during and post product development [1-4]. One of the most commonly used methods to elicit patient preferences are discrete choice experiments (DCE) [5, 6]. DCEs are based on Random Utility Theory (RUT) [7]. In DCEs, respondents complete multiple choice tasks in which two or more treatment alternatives (described using specific attributes with varying in levels) are presented and the respondent needs to choose the treatment alternative which represents the greatest amount of utility for them [8-11]. The relative importance of these attributes and their individual levels can then be calculated using econometric models [12].

Fundamental to the validity of DCE outcomes is that respondents are informed about the health choice context and the included attributes and attribute levels as intended by the researchers [13, 14]. Good research practices suggest that an educational information section should be included prior to preference elicitation which presents the overall context of the study, describes the attributes and levels, and instructs the respondent on how to complete the choice task [14-16]. This promotes a uniformity in understanding of the context and alternatives presented in a choice task [17] which is needed especially in contexts where there is no interviewer input during the DCE [18, 19]. If patients are not informed or there is no uniformity in understanding, then the preference outcomes generated in these studies may not be comparable and cannot be said to accurately reflect the informed preferences of the respondents, threatening their use and application [14, 17].

Little guidance regarding how educational information should be delivered to participants completing a DCE has been given to date, so researchers often use a simple educational text prior to the choice tasks [20, 21]. However, low levels of health literacy have been identified as a major hurdle in the understanding of text-based health information [22, 23], thus the traditional text format may not be suitable for a significant part of the population [22-26]. In order to overcome this hurdle, the use of educational videos has been proposed as a means to educate those with low levels of health literacy [27-30].

Recent studies have compared preference outcomes when information is presented in standard text vs. non-text formats with mixed results [31-33]. They found that using video based educational material was associated with higher levels of information comprehension [31], consistency in choices [32] or differences in preferences [33, 34]. Further, presenting information in video format instead of text has been found to make completing choice tasks easier for respondents [35].These results imply that presenting educational material in non-text formats may have an impact on patient preferences, but questions remain about the generalizability of these findings and whether this impact can be replicated in multiple populations.

This study aimed to assess 1) the impact that educating patients using video-based educational materials with a voiceover has on patient preferences compared to traditional text and 2) whether this impact is consistent between two countries.

2. METHODS

A case study assessing diabetes patient preferences for glucose monitoring technology in self-monitoring of blood glucose (SMBG) was used for this study. This case study was identified as suitable for a study of educational materials as poor diabetes knowledge and low levels of health literacy are often reported in diabetes patient populations and are often found to be associated with poor glycemic control [36-43]. The need for education in these areas was further highlighted by clinical experts and patient representatives during the development of the DCE who reported concerns about whether patients were informed enough about diabetes complications and the need for SMBG to make an informed decision.

2.1 Study Sample and Ethics

Respondents were recruited in the Netherlands (NL) and Poland (PO) through an online panel provider (SurveyEngine) from January to March 2020. The Netherlands and Poland were chosen as examples of 'Western' and 'Eastern' European countries with partial and no reimbursement for glucose monitoring devices at the time of data collection (respectively) [44-46], allowing for the inclusion of a cost attribute in the DCE. Inclusion criteria for the study were self-reported diagnosis of diabetes (Type 1, Type 2, or Other), age ≥18, residing in the Netherlands or Poland, able to read and understand Dutch or Polish, and with access to an internet-connected laptop or computer. This study was approved by the Medical Research Ethics Committee of the UMC Utrecht (WAG/mb/19/045208) and was conducted according to the principles of the Declaration of Helsinki. All respondents provided written informed consent prior to participating in the study.

2.2 DCE Development

The DCE attributes and levels were developed according to best practices in a threefold process. First, a scoping literature review was done to identify relevant attributes of glucose-monitoring technology from previously published studies. Results of this review were used to develop an interview guide for use in interviews with diabetes patients in NL (N=9), a focus group with patients in PO (N=10), as well as interviews with clinical diabetes experts (n=5), patient organization representatives in NL and PO (n=2), and pharmaceutical industry representatives involved in glucose monitoring device development (n=4). The resulting list of 12 potentially relevant attributes were discussed by the research team and reduced based on relevance, completeness, non-redundancy, operationality, and preferential independency to a final list of 7 attributes with 2 to 4 levels (see Table 1). The information on the attributes and levels were developed by the researchers according to best practices [15, 16]. One attribute ('Monthly costs') was standardized between the two countries using purchasing power parity weights to assure that the relative value of the levels was similar given the differences in wealth between the two countries [47].

Table 1: Attributes and levels for the discrete choice experiment

Attributes	Level 1	Level 2	Level 3	Level 4
Precision compared to finger-pricking [a]	Less accurate than finger-pricking (higher or lower by 0.6 mmol/L (*10.8 mg/dL)	Less accurate than finger-pricking (higher or lower by 0.3 (*5.4 mg/dL)	Accurate as finger-pricking §	---
Average number of finger-pricks per day [b]	4 §	2	0	---
Effort to check [c]	High effort: you need to measure your glucose levels yourself	Moderate effort: you scan a sensor to check glucose levels	Low effort: glucose levels automatically sent to you §	---
Probability of getting skin irritation or redness [d]	35% chance of skin irritation or redness	20% chance of skin irritation or redness	5% chance of skin irritation or redness	No chance of skin irritation or redness §
Monthly costs [e]	€250 (*550zl)	€175 (*390zl)	€100 (*220zl)	€25 (*55zl) §
Glucose information [f]	Current Glucose level §	Current Glucose level and arrow	Current Glucose level and a graphic of your level trends over the day	---
Alarms [g]	No §	Yes	---	---

* Unit equivalents shown for Polish survey

§ Reference level

(a-g) Attribute explanations as presented to patients:

A. Some glucose monitors are more precise than others. Finger-pricking is generally regarded as the most accurate way to measure glucose levels. Measurements from devices that use sensors can be just as accurate, but can also be less accurate than finger-pricking, especially if your glucose levels are very high or very low. For example, if your glucose level is 6 mmol/L and you measure it with a device that is off by 0.6 mmol/L, then this device can say your glucose is anywhere from 5.4 to 6.6 mmol/L

B. This is how many times you would need to do a finger-prick-test each day on an average day. This number could be higher on days when you feel the need to test more often like when you're sick, but we want you to picture an average day. Sometimes, this is your only method of measuring your glucose levels. Or, you might need to do finger-prick-tests to confirm the levels from another device

C. This means how much effort you need to give to check your blood glucose levels. High effort checking means you need to stop what you're doing and concentrate on measuring your levels. You need to wash your hands, get out your device equipment, prick your finger, put blood on a strip, check the results, and then clean everything up. Moderate effort checking means you need to get out a small device and use it to scan the sensor on your body to obtain your glucose levels. Low effort checking means your glucose levels are automatically sent to a device which you can view at any time. This could be a dedicated glucose device, your phone, or a smartwatch. You don't need to do anything to have your blood glucose levels sent through, just look at the device to check.

D. A chance of skin irritation or redness around a sensor means a redness or itchy rash on the skin around or under the sensor. This is similar to having an itchy allergic reaction and can be rather uncomfortable or irritating. The sensor will need to be removed and replaced in a different spot. This skin irritation and redness usually lasts until after the sensor is replaced. Not all sensor have this side effect so chances of getting the side effect can differ per device. If a device gives you a 15% chance, this means that 15 out of a 100 people who get this device experience skin irritation and redness while 85 out of a 100 people do not experience this.

E. This means how much money you need to pay out-of-pocket per month in order to check your blood glucose. Please note that this is money that is not reimbursed by your insurance. This could be money needed to pay for devices, sensors, or strips used.

F. This means how your glucose levels are presented to you. This information could be only your current glucose level (you only see a digital number like 8.3 mmol/L). This could be your current glucose level with an arrow showing how your blood glucose is changing as compared to your previous measurement (increasing, decreasing, stable). Or, it could show your current glucose level with a graphic of your blood glucose levels over the day.

G. Your device will give you a beeping alarm (like a phone notification) any time your blood glucose levels are (getting) too high or too low.

2.2.1 DCE Design

A Bayesian D-efficient design with three blocks of 12 choice tasks was developed using NGene 1.0 software. Initial priors were derived using best guess estimates based on the available literature, interviews, and researcher knowledge. In each choice task patients were instructed to imagine that their doctor told them to check their blood glucose levels at least four times per day, and gave them device options to choose from. Respondents had to choose between two hypothetical glucose monitor alternatives. After this, they were offered the opportunity to use the alternative option or opt-out and use a standard finger-prick test for their care (see Figure 1) [48, 49]. This design ensured that respondents reviewed the choice alternatives while still reflecting realistic choice scenarios as patients can always choose a standard finger-prick test for SMBG in real-life. Each hypothetical alternative was described using the seven attributes with varying levels. The finger-prick test was described using the same seven attributes as the two glucose monitoring devices (see Figure 1). Additional information explaining the attributes and levels (including an icon array) was available in a pop-up window which the participant could view during the task by hovering over the attribute. Respondents were given two practice DCE choice-tasks before the main exercise started as a way to familiarize themselves with the task. The DCE design was updated prior to final data collection after a pilot-test of N=99 NL patients.

2.3 Educational Material Development

The educational material was created using a framework to help researchers develop educational material for preference studies from the PREFER project [50]. The framework consists of a three-step approach [51]. In the first step, the educational needs of the patient population for understanding the choice context and the choice task were identified. Areas considered in this step include aspects of the study population (e.g., age, disease experience, literacy levels), the disease and treatment context (e.g., impact of disease on work/family/social life, disease knowledge), and the preference task (e.g., preference method used, task complexity). The results of this step can be found in the supplementary information.

Information on diabetes outcomes, the use of glucose monitors in diabetes self-management, and attributes used in choice task were identified as being relevant information for patients to understand in order to ensure the preferences were informed.

In step 2 the content of the educational material was developed. The content consisted of general information related to uncontrolled blood glucose levels in diabetes, short- and long-term diabetes complications associated with hyper- and hypoglycemia, and self-management activities including SMBG. Text related to these topics was extracted from clinical patient information sources and patient websites from five countries (The Netherlands, Poland, United Kingdom, United States, and Australia), as well as printed educational material available through primary and secondary care facilities and patient organizations in the Netherlands.

Figure 1. Example DCE Choice task

Imagine that your doctor told you to check your blood glucose levels at least four times per day. To do this, the doctor offers you different hypothetical devices to choose from.

	Device A		Device B
Precision compared to fingerpricking	Less accurate than fingerpricking (higher or lower by 0.3)		Less accurate than fingerpricking (higher or lower by 0.6)
Average number of fingerpricks per day	0		0
Effort to check	Low effort		Moderate effort
Probability of getting skin irritation or redness	5% chance of skin irritation or redness (5 out of 100)		35% chance of skin irritation or redness (35 out of 100)
Glucose information	Current Glucose level		Current Glucose level and arrow
Alarms	Yes		No
Monthly costs	€25		€175
I prefer:	❍		❍

If you have to choose between the device you have chosen above and the traditional fingerprick-test to check your glucose levels, which one would you prefer? (Please note that a fingerprick-test should be done four times a day, requires high effort to check, does not result in skin irritation or redness, will show your glucose levels, doesn't have an alarm and costs €25 per month).

Select only one answer

❍ **I prefer the device I have selected above**	❍ **I prefer the fingerprick-test**

This text was used as the basis for a narrative script which illustrated the actions and potential outcomes of diabetes self-management in regards to blood glucose levels and diabetes complications. Attributes and level information was developed by the researchers according to best practices [15, 16]. The text took approximately 11 minutes to read and was written at an 8th grade reading level with a Flesch-Kincaid Grade level of 5.8 and a Flesch Reading Ease score of 69.8 (see supplementary information A) [52, 53]. The content was developed first in English and then translated into Dutch (by native language speaking members of the research

team) and Polish (by a professional translation service) using forward-backward translation to ensure comparability of the educational materials between the countries.

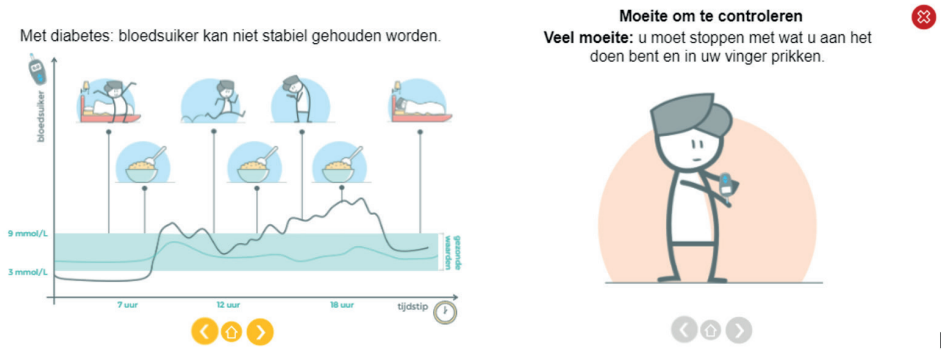
In step 3, the features of the educational material were selected based on the information from step 1. Possible features assessed in this step include levels of realism, simulation, immersion, interactivity, and narrative structure. Based on the high impact of diabetes on patients, their high levels of experience with disease and treatment, and the complexity of managing diabetes, educational materials with high levels of narration, realism, and interactivity was recommended.

Based on this framework, a video consisting of an animated storyline with voiceover was used as it could better present an experienced patient population with realistic narratives and prevent potential literacy issues than other formats and features [27, 54-57]. The animation was developed by MindBytes, a Belgian company with experience in developing educational material for patients in clinical and preference elicitation settings (http://www.mindbytes.be) (see Figure 2, supplementary material for video file). Voiceovers of the text were recorded by native language professional voice actors in the Netherlands and Poland. Risk information for the skin irritation attribute was presented using icon arrays [58]. Respondents were required to access all of the educational material at least once prior to entering the survey. The content of the educational materials was reviewed by a clinical expert (n=1) and pre-tested with patients (n=5) in think-aloud interviews. The Dutch version of the video material lasted 9:52 minutes and the Polish version lasted 9:11 minutes. The educational material was well received in pre-testing. Minor revisions were made based on patient suggestions in regards to the usability of the material.

2.4 Additional Background and Demographic Questions

Respondents completed sociodemographic questions, questions about diabetes type and history, use of medications, and questions related to their current diabetes self-care regimen. Additionally, two brief measures assessing subjective numeracy (the Shortened Subjective Numeracy Scale (SNS-3)) [59] and health literacy (Brief Health Literacy Screener (Chew 3-Items)) [60] were included in the survey. The SNS-3 assesses subjective health numeracy on a 6-point Likert-type scale. On a scale range of 3 to 18, a cutoff of 9.83 was used as a population mean level for high (above cutoff) and low (below cutoff) numeracy. The Chew 3-Items assesses health literacy by asking how often patients need help with medical information material. Health literacy is labeled as inadequate if patients report needing help 'sometimes' or 'more often' in one of these three questions. Finally, respondents were asked to rate the length of the survey, ease of understanding the educational material, and ease of understanding and ease of completing the choice tasks based on a 6-point Likert scale.

Figure 2. Example images from the animated storyline showing general diabetes information (left translated from Dutch "With diabetes: blood sugar cannot be kept stable") and attribute information (right translated from Dutch; "Effort to check - High effort: You need to stop what you are doing and prick your finger.")



2.5 Survey Design

Respondents were randomized to arms receiving the educational material in either the standard text format or in the video format (hereafter referred to as NL-T, NL-V, PO-T, PO-V for Dutch/Polish, Text/Video information samples respectively). All data was collected in a one-time, online survey.

2.6 Statistical analysis

All analyses were conducted using STATA 14 with the clogit and mixlogit package. Respondents were removed from analysis if the person self-reported as not having diabetes or if the amount of time used to complete the survey was less than 70% of the mean response time for their country and educational material arm. Effects coding was used in which the reference category was coded as -1 for all attributes [61]. A Swait and Louviere test was conducted to identify scale differences across respondents from the two countries and see if the datasets could be combined [62, 63]. Significant differences in parameter estimates were found between the countries (χ^2 (1) = 30.296, $p < 0.001$), thus data from the Netherlands and Poland were analyzed separately. A Swait and Louviere test was also conducted to identify potential scale differences between educational tool arms within each country.

Preference estimates for each educational material arm were assessed using a mixed-effects logistic regression with random effects and a normal distribution to account for heterogeneity of preferences which often exists within patient populations [64]. For robust results, each model was built using 14,000 Halton draws [65]. The following utility model was used:

$$U_{Aci} = \beta_0 + \beta_{1i} * precision_{0.3} + \beta_{2i} * precision_{0.6} + \beta_{3i} * pricks\ per\ day_{2x} + \beta_{4i} * effort_{moderate} + \beta_{5i} * skin\ irritation_{20\%} + \beta_{6i} * skin\ irritation_{35\%} + \beta_{7i} * monthly\ costs_{€100} + \beta_{8i} * monthly\ costs_{€175} + \beta_{9i} * monthly\ costs_{€250} + \beta_{10i} * information_{arrow} + \beta_{11i} * information_{trendline} + \beta_{12i} * alarms_{none} + \varepsilon$$

$$U_{Bci} = \beta_{1i} * precision_{0.3} + \beta_{2i} * precision_{0.6} + \beta_{3i} * pricks\ per\ day_{2x} + \beta_{4i} * effort_{moderate} + \beta_{5i} * skin\ irritation_{20\%} + \beta_{6i} * skin\ irritation_{35\%} + \beta_{7i} * monthly\ costs_{€100} + \beta_{8i} * monthly\ costs_{€175} + \beta_{9i} * monthly\ costs_{€250} + \beta_{10i} * information_{arrow} + \beta_{11i} * information_{trendline} + \beta_{12i} * alarms_{none} + \varepsilon$$

$$U_{\text{Fingerprick } ci} = \beta_{13i} + \varepsilon$$

In this model, the utility (U) of an alternative (A, B, or finger-prick) in a specific decision context (c) for an individual (i) was derived as the sum of the attribute-level estimates indicating the relative importance of each attribute level ($\beta_1 - \beta_{12}$) or the relative importance of the status quo finger-prick glucose monitor (β_{13}). B_0 represents a constant term reflecting a left-right bias when choosing an alternative (i.e., choosing the left alternative when the coefficient is significant and has a positive sign, right if significant and negative). Stochastic factors for this alternative are included in the utility function as a random error term ε.

The preference estimates from the mixed-logit model were used to calculate attribute relative importance scores (RIS) by identifying the attribute with the greatest absolute difference between most and least valued level and using this as the reference [66]. The RIS for each attribute is the quotient of the absolute difference of the most and least valued level of that attribute and the reference value. This results in a normalized scale with the attribute with the greatest difference in level values assigned 1 and all other attributes valued proportionally to it. Feedback on the educational material and meta-data related to time spent in the educational material and time spent in the survey was compared between the educational material arms using t-tests or χ^2 tests where applicable. A significance level of 0.05 was used for all analyses.

3. RESULTS

3.1 Sample characteristics

In total, 981 completed surveys were analyzed after n=56 responses were removed for completing the survey too quickly. The Polish responses (n=522) were evenly split between educational material formats (PO-V n=261, PO-T n=261). The Dutch responses (n=459) had slightly more video responses (NL-V n=233) than standard text response (NL-T n= 226). The sample demographics can be found in Table 2. Significant differences were found between the countries with the Polish sample being younger, with fewer years of diabetes, and higher levels of education and health numeracy. Significant differences were found within the Netherlands between the two educational information formats for type of diabetes (χ^2 (2) = 8.19, p = .017). Both survey populations were more highly educated with a higher prevalence of Type

1 diabetes than would be expected from general diabetes population characteristics and the Polish sample was younger than the general diabetes population [67, 68].

Table 2. Sample demographics

Characteristics	Dutch respondents		Polish respondents	
	N= 226 Text	N=233 Video	N=261 Text	N=261 Video
Age in years *(mean ± sd)	51.6 ± 17.2	50.5 ± 17.8	39.4 ± 13.4	39.0 ± 13.1
Sex (n, %)				
Females	116 (51.3)	115 (49.4)	125 (47.9)	131 (50.2)
Males	110 (48.7)	117 (50.2)	134 (51.3)	130 (49.8)
Other	0 (0.0)	1 (0.4)	2 (0.8)	0 (0.0)
Type of diabetes (n, %)Ŧ				
Type 1	65 (28.8)	59 (25.3)	83 (31.8)	84 (32.2)
Type 2	158 (69.9)	159 (68.2)	167 (64.0)	164 (62.8)
Other	3 (1.3)	15 (6.4)	11 (4.2)	13 (5.0)
Glucose monitor type* (n, %)				
CGM or FGM	38 (16.8)	46 (19.7)	39 (14.9)	33 (12.6)
Finger-prick testing only	128 (56.6)	122 (52.4)	211 (80.8)	215 (82.4)
None	60 (26.5)	65 (27.9)	11 (4.2)	13 (5.0)
Number of years having * diabetes (mean ± sd, (median, range))	9.5 ± 9.1 (6.5, 0-60)	10.4 ± 9.4 (8, 0-46)	6.1 ± 7.1 (3, 0-53)	6.3 ± 6.9 (4, 0-50)
Health literacy (n, %)				
Adequate	102 (45.1)	103 (44.2)	113 (43.3)	111 (42.5)
Inadequate	124 (54.9)	130 (55.8)	148 (56.7)	150 (57.5)
Health numeracy* (n, %)				
High	195 (86.3)	194 (83.3)	243 (93.1)	241 (92.3)
Low	31 (13.7)	39 (16.7)	18 (6.9)	20 (7.7)
Educational level ¥ (n, %)				
Tertiary	100 (44.3)	84 (36.1)	134 (51.3)	154 (59.0)
Upper-Secondary/Vocational	114 (50.4)	130 (55.8)	127 (48.7)	107 (41.0)
Secondary or Lower	12 (5.3)	19 (8.2)	(0.0)	0 (0.0)
Dominant Cost decision making (n, %) ***	73 (32.3)	64 (27.5)	57 (21.8)	48 (18.4)

Significant differences between countries at * p<0.05 ** p<0.01 *** p<0.001
Ŧ Significant differences between educational tool formats in the Dutch sample at p<0.05;
¥ Education levels were based on the definitions used by the OECD.

3.2 Main effects model

The preference estimates of the mixed-effects models can be found in Table 3. A significant left-right bias was found in all arms with respondents tending to choose the left option. For all groups, the coefficients followed logical patterns where "more attractive" levels were valued higher than "less attractive". All attribute-level estimates were significant for at least one level in one of the arms. No significant differences in preferences or scale parameters were found between the text and video educational material arms using the Swait and Louviere test in either NL (χ^2 (1, 459) =2.492, p = 0.114) or PO (χ^2 (1, 522) =1.600, p = 0.206). Dominant decision-making behavior was found for the lowest cost level in all educational material arms (see table 2), however there was no significant difference in the amount of dominant decision-making between the educational arms within countries (NL: χ^2 (1, 459) = 1.280, p = 0. .258; PO: χ^2 (1, 522) =1.600, p = 0.206). These respondents were not removed for the final analysis as Costs are a major concern of the general patient population and the results need to be understood in relation to this attribute. Preference estimates of mixed-effects models excluding these respondents can be found in supplementary material table 3A.

For the Dutch sample, *Costs* were the most important attribute. The only attribute which was not significant for the Dutch sample was *Glucose Information*. For the Dutch sample, significant heterogeneity was found in preferences for all attributes and the status quo finger-prick-test except for *Effort to Check*, *Alarm* and *Glucose Information*. Differences in heterogeneity of preferences were found between the educational material arms. Heterogeneity was found for a moderate *Chance of Skin Irritation* and the € 175 cost level only in the NL-T arm, while heterogeneity of preferences for the € 250 cost levels was only found in the NL-V arm.

For the Polish sample, *Costs* were similarly the most important attribute, while improving *Effort to Check* was not significant for the Polish sample. Significant heterogeneity of preferences was found for all attributes and the status quo finger-prick-test except for *Chance of Skin Irritation*. Differences in heterogeneity of preferences was only found for the attribute level of *Glucose Information* shown as a daily trend. Here the NL-V arm had significant heterogeneity of preferences, but the PO-T arm did not. Finally, the NL-V arm had a much lower valuation of the status quo finger-prick-test compared to all other arms.

3.3 Relative Importance Scores

The RIS for all samples can be seen in Figure 3. *Costs* were 2.7 to 4.3 times more important than the second most important attribute and 15.2 to 56.1 times more important than the least important attribute across the arms. After *Costs*, for all the arms the attributes could be separated into two tiers with *Precision*, *Number of finger-pricks*, and *Chance of Skin irritation* being the next most important attributes, and *Alarm*, *Glucose information*, and *Effort to check* being the least important attributes. Differences were observed in RISs, but they were relatively small within the tiers (.02-.23 in the top tier and .03-.06 in the lower tier). The exception to this was *Precision* for the NL-V arm which was clearly more important than the other in-tier attributes.

Table 3. Attribute-level estimates for the mixed-logit model

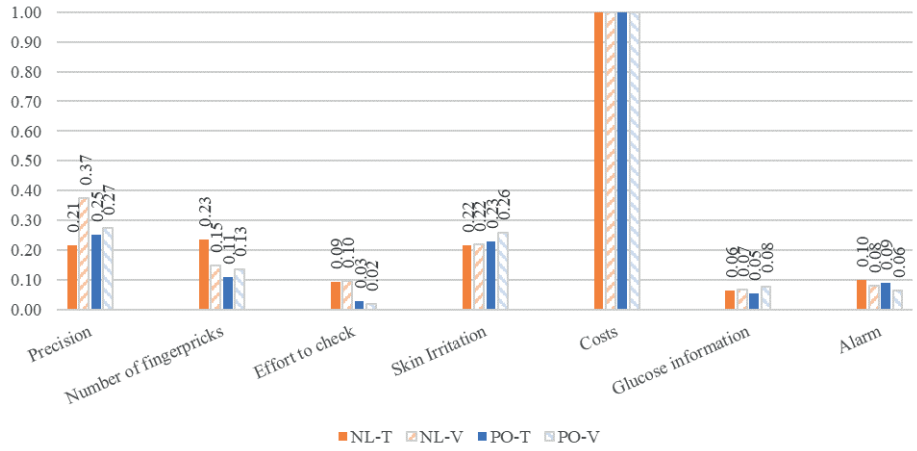
		Netherlands						Poland					
		Text			Video			Text			Video		
Attribute		**Beta**	**SE**	**p>|z|**	**Beta**	**SE**	**p>|z|**	**Beta**	**SE**	**p>|z|**	**Beta**	**SE**	**p>|z|**
Precision													
High §		0.32			0.57			0.47			0.45		
Medium	Mean	0.02	0.06	-	0.03	0.06	-	-0.09	0.05	-	-0.10	0.05	*
	SD	-	-	-	-	-	-	-	-	-	-	-	-
Low	Mean	-0.34	0.08	***	-0.60	0.08	***	-0.39	0.07	***	-0.35	0.06	***
	SD	0.55	0.10	***	-0.65	0.09	***	0.60	0.07	***	0.52	0.08	***
Number of Finger-pricks per day													
0 §		0.36			0.23			0.17			0.13		
2	Mean	-0.36	0.06	***	-0.23	0.05	***	-0.17	0.04	***	-0.13	0.04	**
	SD	-0.46	0.07	***	0.38	0.06	***	0.35	0.05	***	0.35	0.06	***
Effort to Check													
Low §		0.14			0.15			0.05			0.02		
Moderate	Mean	-0.14	0.04	***	-0.15	0.04	***	-0.05	0.03	-	-0.02	0.03	-
	SD	-	-	-	-	-	-	-	-	-	-	-	-
Chance of Skin Irritation													
Low §		0.36			0.40			0.37			0.35		
Moderate	Mean	-0.06	0.07	-	-0.12	0.06	-	0.02	0.06	-	0.00	0.06	-
	SD	-0.20	0.08	*	-	-	-	-	-	-	-	-	-
High	Mean	-0.30	0.07	***	-0.28	0.06	***	-0.39	0.06	***	-0.35	0.06	***
	SD	-	-	-	-	-	-	-	-	-	-	-	-

(cont'd)

		Netherlands						Poland					
		Text			Video			Text			Video		
Attribute		**Beta**	**SE**	**p>|z|**	**Beta**	**SE**	**p>|z|**	**Beta**	**SE**	**p>|z|**	**Beta**	**SE**	**p>|z|**
Costs													
€25 §		1.47			1.47			1.41			1.01		
€ 100	Mean	0.32	0.08	***	0.23	0.07	**	0.42	0.07	***	0.25	0.07	***
	SD	1.57	0.14	***	1.63	0.16	***	2.12	0.15	***	1.67	0.13	***
€ 175	Mean	-0.19	0.08	*	-0.04	0.07	-	-0.02	0.07	-	0.12	0.08	-
	SD	-	-	-	-	-	-	-0.28	0.12	*	-0.71	0.09	***
€ 250	Mean	-1.60	0.16	***	-1.65	0.15	***	-1.81	0.14	***	-1.38	0.13	***
	SD	-	-	-	0.19	0.09	0.031	-0.32	0.09	***	-0.36	0.09	***
Glucose Information													
Only current glucose value §		-0.12			-0.09			-0.07			-0.07		
Current glucose value with arrow	Mean	0.04	0.06	-	-0.02	0.06	-	0.01	0.05	-	-0.04	0.05	-
	SD	-	-	-	-	-	-	-0.01	0.08	-	-0.08	0.10	-
Current glucose value with day trend	Mean	0.08	0.06	-	0.11	0.06	-	0.06	0.05	-	0.11	0.05	*
	SD	-	-	-	-	-	-	-	-	-	-0.16	0.07	*
Alarm													
Yes §		0.15			0.12			0.15			0.08		
No	Mean	-0.15	0.04	***	-0.12	0.04	***	-0.15	0.03	***	-0.08	0.03	*
	SD	-	-	-	-	-	-				-0.22	0.05	***
Left-Right bias		0.40	0.08	***	0.35	0.08	***	0.42	0.07	***	0.27	0.07	***
Status quo finger-prick	Mean	0.63	0.28	0.021	0.26	0.28	0.360	0.23	0.21	0.289	-1.42	0.25	***
	SD	4.56	0.42	***	4.60	0.37	***	4.39	0.41	***	4.56	0.35	***

§ Ref: reference;
SD: Standard Deviation (only shown when SD was found to be significant)
* indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$; - indicates not significant

Figure 3. Attribute Relative Importance Scores by Country and Educational Tool Format



Note: Attribute were standardized to 1 based on the attribute with the largest difference between highest and lowest levels which was cost in all arms (difference between highest and lowest cost level by arm: NL-T= 3.07, NL-V=3.12, PO-T=3.39, NL-V=2.85)

3.4 User Experience and Meta-data analysis

Respondent feedback and mean times to complete the educational material can be found in Table 4. Respondent feedback on ease of understanding the educational material was more positive for those who saw the video educational material, although the difference was small. No significant differences were found in reported ease of understanding or completing the DCE. The mean time in educational material ranged from 2.36 to 3.84 minutes. Approximately 20.5% of the text respondents reached the end of the text information in under 30 seconds and 22.6% of video respondents reached the end of the video in under 30 seconds. Significant differences in time spent in the educational material were found for different levels of education for the NL-T arm, and level of health literacy for the NL-T, NL-V, and PO-T arms. A regression analysis on the amount of time spent in the tool found that different patient characteristics were associated with the total time (see supplementary table 3B). For both the PO-T and PO-V arms, being older and female was associated with increased time in the educational materials. For the PO-V arm, CGM use was associated with decreased time in the educational materials. For the NL-T arm, being older and female with more years of having diabetes was associated with increased time in the educational materials. For the NL-V arm, being older and having higher health literacy was associated with increased time in the educational materials, but having T1DM and using a FP or CGM for SMBG were associated with decreased time in the educational materials.

4. DISCUSSION

This was a first of kind study to assess the impact of educational material format in a randomized study in multiple countries. We developed educational materials using an evidence-based framework to address potential gaps in knowledge related to general diabetes information and the use of glucose monitors as a part of diabetes self-management. While some preference differences were found between the arms, no interpretable pattern of differences in the relative importance of attributes could be identified among respondents receiving either the text or video educational material.

Table 4. User Feedback and Meta-data regarding time spent in education material by education level, health literacy level, and health numeracy level M(SD)

	Netherlands		**Poland**	
	Text	**Video**	**Text**	**Video**
Respondent Feedback				
Length of Survey Mean (SD)	4.02 (0.99)	4.01 (0.99)	3.93 (0.99)*	3.77 (0.96)*
Ease of understanding educational material Mean (SD)	4.63 (1.32)***	5.02 (0.98)***	5.011 (1.10)**	5.29 (1.03)**
Ease of understanding DCE Mean (SD)	4.75 (1.38)	4.67 (1.38)	5.00 (1.26)	4.97 (1.23)
Ease of answering DCE Mean (SD)	4.64 (1.39)	4.57 (1.33)	4.84 (1.29)	4.84 (1.29)
Time in minutes Mean (SD, range)	2.41 (2.53, 0.20-23.46)	2.93 (3.32, 0.34-12.79)	2.36 (2.36, 0.20-14.99)	3.84 (4.36, 0.8-26.91)
Educational level				
Tertiary	2.16 (3.04)*	3.33 (3.63)	2.15 (2.27)	4.28 (4.98)
Upper-Secondary/ Vocational	2.43 (1.89)*	2.71 (3.13)	2.47 (2.41)	3.65 (4.06)
Secondary or Lower	4.35 (2.57)*	2.62 (3.15)	-	-
Health literacy				
Adequate	2.11 (2.11)*	2.44 (3.11)*	1.99 (2.19)**	3.40 (3.94)
Inadequate	2.78 (2.93)*	3.54 (3.48)*	2.85 (2.49)**	4.47 (4.81)
Health numeracy[T]				
High	2.78 (2.45)	2.96 (3.06)	1.99 (1.52)	3.70 (3.73)
Low	2.36 (2.54)	2.92 (3.38)	2.39 (2.41)	3.85 (4.41)

T: The mean sample score was used for the cut-off between high and low groups for the health numeracy measure, SNS-3; * indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$; S.D. = standard deviation

These findings replicate to a certain extent those found by Bywall et al. and Lim et al., but not those found by Vass et al. [31-33]. Both Bywall et al. and Lim et al. found differences between samples when presented with either video or text-based information. In the current study, differences between samples were identified when comparing the preference estimates and RISs, but these differences were not statistically significant within each country. Lim et al. supported their findings through evidence of patients being more informed, but the practical differences in knowledge were relatively small. Thus, the question is whether differences found were related to more informed patients or to heterogeneity of patient preferences often found when comparing patient populations [69-73].

One major issue explaining the lack of differences in the current study is that respondents did not appear to attend to all the content of the educational materials thus limiting the exposure needed to cause an effect. Non-attendance to the information is evidenced in the short amount of time that respondents spent on the education material regardless of the format (2.36-3.84 minutes, while the video took over 9 minutes to view entirely, and reading the text-based information was expected to take 11 minutes). The long duration of the educational material may have led to non-attendance, but previous research with educational materials of comparable length has found that patients engage with the educational material when the treatment context was novel to the respondents [31, 74]. The non-attendance in our study could be related to the high levels of disease experience that the sample had. For the Dutch respondents experience with SMBG and more years with diabetes was related to less time in the educational material, but this was not true for the Polish population. Another possible explanation is that the respondents viewed the material as being too general for their needs as it covered a wide amount of information. Previous research has found that patients value information more when it is tailored to their individual informational needs and patients may ignore information that they view as not being relevant to them [75-78]. Unfortunately, years of experience with managing diabetes does not necessarily mean that patients are sufficiently educated on the importance of adequate diabetes self-management [42]. This may especially hold for those patients with low levels of health literacy [41-43]. The identified need from clinical experts and patient representatives to inform respondents highlights the difficulties that future researchers may encounter to ensure that respondents who need additional information actually engage with the educational material. Clinical experts and patient representatives may have more complete understanding of the relevant information needed for diabetes care and objective insights into the educational needs of patients. However, patients may not recognize that there is a need for additional education and thus view the information provided as not being relevant for them. There is a need for future research to identify ways to communicate that there is a need for further education and that the educational material is relevant to them in order to increase engagement with the educational material.

One possible way is through tailoring of information. Tailoring of information to the specific needs of the patients would make the information more relevant, which has previously been found to increase engagement with material [75-78]. Aspects such as patient demographics or knowledge questions could be used to tailor the information to only what is relevant to the individual patient, reducing the burden to patients and increasing engagement with educational materials. Tailoring of information could also identify the formats that the individual patients respond best to in heterogeneous populations, ensuring that respondents can properly understand the information presented to them. Increasing engagement with educational materials in patient preference studies should be a topic of future research as simply presenting information in a video format does not necessarily mean patients will access it.

4.1 Strengths and Limitations

Despite the study being a multi-country, randomized case study, there are a few study limitations which prevent better understanding of how patients interacted with the educational material. First, the video was embedded on one webpage thus the start and stop times were based on the time that a respondent entered and left the page. This limits our ability to see whether there were specific points in the video which patients were more engaged and spent more or less time on. Second, patients were recruited via an online panel where respondents register themselves so it is likely that respondents possess at least a basic level of health and digital literacy, which may not be representative of the entire diabetes population. Hence, it is unclear if the video would have been of more use in a sample of diabetes patients recruited through clinical channels, who may have more difficulty with text or numeric health information and for whom educational information presented in a video format is likely more suitable [27, 54-57]. These panel respondents are also compensated based on completion of the survey so they may have prioritized quick completion. Further, our patient population had higher levels of education than what you would expect from the general population [68, 79, 80]. Finally, we did not include an objective measure of diabetes knowledge which would have allowed us to analyze whether the patients were well-informed prior to participating in the preference study. Patients could have already been well-informed on the information presented making the educational material unnecessary.

5. CONCLUSION

Educating respondents prior to a preference study remains a priority for stakeholders who use patient preference information. Simply providing educational material in a video with animations and voiceovers did not result in significant differences in preference outcomes when compared to text. This may have been due to a lack of engagement with the educational materials in our study, as well as to a high level of familiarity with the topic of the study in this group of experienced patients. Future research should look at ways to increase engagement with educational materials, as well as ways to better tailor information to the needs of individual patients participating in preference studies.

REFERENCES

1. Whichello, C., et al., Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison. BMC Med Inform Decis Mak, 2020. **20**(1): p. 114.
2. Whichello, C., et al., An overview of critical decision-points in the medical product lifecycle: Where to include patient preference information in the decision-making process? Health Policy, 2020. **124**(12): p. 1325-1332.
3. van Overbeeke, E., et al., Factors and situations influencing the value of patient preference studies along the medical product lifecycle: a literature review. Drug Discov Today, 2019. **24**(1): p. 57-68.
4. Janssens, R., et al., Opportunities and challenges for the inclusion of patient preferences in the medical product life cycle: a systematic review. BMC Med Inform Decis Mak, 2019. **19**(1): p. 189.
5. Soekhai, V., et al., Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discov Today, 2019. **24**(7): p. 1324-1331.
6. Ryan, M., K. Gerard, and M. Amaya-Amaya, Using discrete choice experiments to value health and health care. Vol. 11. 2007: Springer Science & Business Media.
7. Ryan, M., K. Gerard, and M. Amaya-Amaya, Using Discrete Choice Experiments to Value Health and Health Care. The Economics of Non-Market Goods and Resources. 2008, Dordrecht: Springer.
8. Ryan, M., K. Gerard, and M. Amaya-Amaya, Discrete Choice Experiments in a Nutshell, in Using Discrete Choice Experiments to Value Health and Health Care, M. Ryan, K. Gerard, and M. Amaya-Amaya, Editors. 2008, Springer Netherlands: Dordrecht. p. 13-46.
9. Lancaster, K.J., A New Approach to Consumer Theory. Journal of Political Economy, 1966. **74**(2): p. 132-157.
10. Manski, C.F., The structure of random utility models. Theory and Decision, 1977. **8**(3): p. 229-254.
11. McFadden, D., Econometric models of probabilistic choice. Structural analysis of discrete data with econometric applications, 1981. **198272**.
12. Ryan, M. and K. Gerard, Using discrete choice experiments to value health care programmes: current practice and future research reflections. Applied health economics and health policy, 2003. **2**(1): p. 55-64.
13. Pearce, A., et al., Respondent Understanding in Discrete Choice Experiments: A Scoping Review. Patient, 2021. **14**(1): p. 17-53.
14. US Department of Health and Human Resources Food and Drug Administration, Patient Preference Information – Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling: Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. 2016, U. S. Food and Drug Administration: Silver Spring, MD.
15. Bridges, J.F., et al., Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health, 2011. **14**(4): p. 403-13.
16. Lancsar, E. and J. Louviere, Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics, 2008. **26**(8): p. 661-77.
17. Brown, T.C., Introduction to stated preference methods, in A primer on nonmarket valuation. 2003, Springer. p. 99-110.
18. de Bekker-Grob, E.W., M. Ryan, and K. Gerard, Discrete choice experiments in health economics: a review of the literature. Health Econ, 2012. **21**(2): p. 145-72.

19. Soekhai, V., et al., Discrete Choice Experiments in Health Economics: Past, Present and Future. Pharmacoeconomics, 2019. **37**(2): p. 201-226.

20. Smith, I.P., et al., Methodological Priorities for Patient Preferences Research: Stakeholder Input to the PREFER Public-Private Project. Patient, 2021. **14**(5): p. 449-453.

21. Louviere, J.J. and E. Lancsar, Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law, 2009. **4**(Pt 4): p. 527-46.

22. Nutbeam, D., The evolving concept of health literacy. Soc Sci Med, 2008. **67**(12): p. 2072-8.

23. Andrus, M.R. and M.T. Roth, Health literacy: a review. Pharmacotherapy, 2002. **22**(3): p. 282-302.

24. Paasche-Orlow, M.K., et al., The prevalence of limited health literacy. J Gen Intern Med, 2005. **20**(2): p. 175-84.

25. Kutner, M., et al., The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. National Center for Education Statistics, 2006.

26. Sorensen, K., et al., Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). Eur J Public Health, 2015. **25**(6): p. 1053-8.

27. Moran, M.B., et al., A pilot test of the acceptability and efficacy of narrative and non-narrative health education materials in a low health literacy population. J Commun Healthc, 2016. **9**(1): p. 40-48.

28. Ferguson, L.A., Implementing a Video Education Program to Improve Health Literacy. The Journal for Nurse Practitioners, 2012. **8**(8): p. e17-e22.

29. Bandyopadhyay, M., et al., Accessibility of web-based health information for women in midlife from culturally and linguistically diverse backgrounds or with low health literacy. Aust N Z J Public Health, 2022. **46**(3): p. 269-274.

30. Hart, T.L., et al., Development of multimedia informational tools for breast cancer patients with low levels of health literacy. Patient Educ Couns, 2015. **98**(3): p. 370-7.

31. Lim, S.L., et al., Are Videos or Text Better for Describing Attributes in Stated-Preference Surveys? Patient, 2020. **13**(4): p. 401-408.

32. Vass, C.M., et al., A Picture is Worth a Thousand Words: The Role of Survey Training Materials in Stated-Preference Studies. Patient, 2020. **13**(2): p. 163-173.

33. Bywall, K.S., et al., Does being exposed to an educational tool influence patient preferences? The influence of an educational tool on patient preferences assessed by a discrete choice experiment. Patient Educ Couns, 2021. **104**(10): p. 2577-2585.

34. Sandorf, E.D., M. Aanesen, and S. Navrud, Valuing unfamiliar and complex environmental goods: A comparison of valuation workshops and internet panel surveys with videos. Ecological Economics, 2016. **129**: p. 50-61.

35. Navrud, S., Communication devices in contingent valuation surveys—Experiments with video, in Determining the Value of Non-Marketed Goods. 1997, Springer. p. 273-298.

36. Phillips, E., R. Rahman, and M. Mattfeldt-Beman, Relationship Between Diabetes Knowledge, Glycemic Control, and Associated Health Conditions. Diabetes Spectr, 2018. **31**(2): p. 196-199.

37. Bukhsh, A., et al., Association of diabetes knowledge with glycemic control and self-care practices among Pakistani people with type 2 diabetes mellitus. Diabetes Metab Syndr Obes, 2019. **12**: p. 1409-1417.

38. Bains, S.S. and L.E. Egede, Associations between health literacy, diabetes knowledge, self-care behaviors, and glycemic control in a low income population with type 2 diabetes. Diabetes Technol Ther, 2011. **13**(3): p. 335-41.

39. Colleran, K.M., B. Starr, and M.R. Burge, Putting diabetes to the test: Analyzing glycemic control based on patients' diabetes knowledge. Diabetes Care, 2003. **26**(7): p. 2220-1.

40. Al-Qazaz, H.K., et al., Diabetes knowledge, medication adherence and glycemic control among patients with type 2 diabetes. International journal of clinical pharmacy, 2011. **33**(6): p. 1028-1035.

41. Marciano, L., A.L. Camerini, and P.J. Schulz, The Role of Health Literacy in Diabetes Knowledge, Self-Care, and Glycemic Control: a Meta-analysis. J Gen Intern Med, 2019. **34**(6): p. 1007-1017.

42. Murata, G.H., et al., Factors affecting diabetes knowledge in Type 2 diabetic veterans. Diabetologia, 2003. **46**(8): p. 1170-8.

43. Zowgar, A.M., M.I. Siddiqui, and K.M. Alattas, Level of diabetes knowledge among adult patients with diabetes using diabetes knowledge test. Saudi Med J, 2018. **39**(2): p. 161-168.

44. Hohendorff, J., et al., Intermittently Scanned Continuous Glucose Monitoring Data of Polish Patients from Real-Life Conditions: More Scanning and Better Glycemic Control Compared to Worldwide Data. Diabetes Technol Ther, 2021. **23**(8): p. 577-585.

45. Głowińska-Olszewska, B., et al., Monthly use of a real-time continuous glucose monitoring system as an educational and motivational tool for poorly controlled type 1 diabetes adolescents. Advances in medical sciences, 2013. **58**(2): p. 344-352.

46. Heinemann, L., et al., Reimbursement for continuous glucose monitoring: a European view. J Diabetes Sci Technol, 2012. **6**(6): p. 1498-502.

47. OECD. OECD (2022), Purchasing power parities (PPP) (indicator). doi: 10.1787/1290ee5a-en (Accessed on 10 February 2022). 2022 [cited 2019 September].

48. Ghijben, P., E. Lancsar, and S. Zavarsek, Preferences for oral anticoagulants in atrial fibrillation: a best-best discrete choice experiment. Pharmacoeconomics, 2014. **32**(11): p. 1115-27.

49. Veldwijk, J., et al., The effect of including an opt-out option in discrete choice experiments. PLoS One, 2014. **9**(11): p. e111805.

50. The PREFER consortium, PREFER Recommendations - Why, when and how to assess and use patient preferences in medical product decision-making. 2022.

51. Verschueren, S., C. Buffel, and G. Vander Stichele, Developing Theory-Driven, Evidence-Based Serious Games for Health: Framework Based on Research Community Insights. JMIR Serious Games, 2019. **7**(2): p. e11565.

52. Flesch, R., A new readability yardstick. J Appl Psychol, 1948. **32**(3): p. 221-33.

53. Kincaid, J.P., et al., Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975, Naval Technical Training Command Millington TN Research Branch.

54. Calderon, J.L., et al., Improving Diabetes Health Literacy by Animation. Diabetes Educ, 2014. **40**(3): p. 361-372.

55. McCarthy, D.M., et al., What did the doctor say? Health literacy and recall of medical instructions. Med Care, 2012. **50**(4): p. 277-82.

56. Meppelink, C.S., et al., The effectiveness of health animations in audiences with different health literacy levels: an experimental study. J Med Internet Res, 2015. **17**(1): p. e11.

57. Sheridan, S.L., et al., Interventions for individuals with low health literacy: a systematic review. J Health Commun, 2011. **16 Suppl 3**(sup3): p. 30-54.

58. Visschers, V.H., et al., Probability information in risk communication: a review of the research literature. Risk Anal, 2009. **29**(2): p. 267-87.

59. Fagerlin, A., et al., Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Med Decis Making, 2007. **27**(5): p. 672-80.

60. Chew, L.D., K.A. Bradley, and E.J. Boyko, Brief questions to identify patients with inadequate health literacy. Fam Med, 2004. **36**(8): p. 588-94.

61. Bech, M. and D. Gyrd-Hansen, Effects coding in discrete choice experiments. Health Econ, 2005. **14**(10): p. 1079-83.

62. Swait, J. and J. Louviere, The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. Journal of Marketing Research, 1993. **30**(3): p. 305-314.

63. Janssen, E., et al., Patient preferences for mitral valve regurgitation treatment: a discrete choice experiment. Postgrad Med, 2022. **134**(2): p. 125-142.

64. Hensher, D.A. and W.H. Greene, The Mixed Logit model: The state of practice. Transportation, 2003. **30**(2): p. 133-176.

65. Train, K.E., Drawing from Densities, in Discrete Choice Methods with Simulation, K.E. Train, Editor. 2001, Cambridge University Press: Cambridge. p. 205-236.

66. Gonzalez, J.M., A Guide to Measuring and Interpreting Attribute Importance. Patient, 2019. **12**(3): p. 287-295.

67. Volksgezondheidenzorg.info. Diabetes mellitus→Cijfers & Context→Huidige situatie. 2021 [cited 2021 19 January]; Available from: https://www.volksgezondheidenzorg.info/onderwerp/diabetes-mellitus/cijfers-context/huidige-situatie#node-prevalentie-diabetes-huisartsenpraktijk-naar-opleiding.

68. Polakowska, M. and W. Piotrowski, Incidence of diabetes in the Polish population. Results of the Multicenter Polish Population Health Status Study – WOBASZ. Polish Archives of Internal Medicine, 2011. **121**(5): p. 156-163.

69. Hole, A.R., Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment. J Health Econ, 2008. **27**(4): p. 1078-1094.

70. Hiligsmann, M., et al., Are Patients' Preferences Transferable Between Countries? A Cross-European Discrete-Choice Experiment to Elicit Patients' Preferences for Osteoporosis Drug Treatment. Value Health, 2014. **17**(7): p. A385.

71. Kievit, W., et al., Taking Patient Heterogeneity and Preferences into Account in Health Technology Assessments. Int J Technol Assess Health Care, 2017. **33**(5): p. 562-569.

72. Ludwig, K., et al., To What Extent Do Patient Preferences Differ From General Population Preferences? Value Health, 2021. **24**(9): p. 1343-1349.

73. Kaltoft, M.K., et al., Addressing preference heterogeneity in public health policy by combining Cluster Analysis and Multi-Criteria Decision Analysis: Proof of Method. Health Econ Rev, 2015. **5**(1): p. 10.

74. van Overbeeke, E., et al., Patient preferences for gene therapy in haemophilia: Results from the PAVING threshold technique survey. Haemophilia, 2021. **27**(6): p. 957-966.

75. Raynor, D.K., et al., A systematic review of quantitative and qualitative research on the role and effectiveness of written information available to patients about individual medicines. Health Technol Assess, 2007. **11**(5): p. iii, 1-160.

76. Young, A., J. Tordoff, and A. Smith, 'What do patients want?' Tailoring medicines information to meet patients' needs. Res Social Adm Pharm, 2017. **13**(6): p. 1186-1190.

77. Dickinson, R., et al., Suits you? A qualitative study exploring preferences regarding the tailoring of consumer medicines information. Int J Pharm Pract, 2013. **21**(4): p. 207-15.

78. Grime, J., et al., The role and value of written information for patients about individual medicines: a systematic review. Health Expect, 2007. **10**(3): p. 286-98.

79. Rijksinstituut voor Volksgezondheid en Milieu. Diabetes mellitus: Leeftijd en geslacht. 2021 [cited 2021 01/19]; Available from: https://www.vzinfo.nl/diabetes-mellitus/leeftijd-en-geslacht.

80. OECD, Education at a Glance: OECD Indicators Poland. 2019.

SUPPLEMENTARY INFORMATION

Table A1. Factors used to identify educational needs of patient populations

Disease Identifier	Question	Answer
Disease	What is the disease impact on work, family and social life?	Medium: Impacts emotional, social, and physical aspects of the individual's life (QoL)
Patient	How much knowledge/experience does the typical patient have?	Some knowledge/experience with disease/treatments: Has some knowledge/experience but will likely need more information
Treatment: Alternatives	How many alternative treatment options are available?	Few (1-2): There are only 1 or 2 other treatment options available to the patients, easy decision-process
Treatment: Complexity	How complex is the treatment (e.g., administration, dosing, risk-benefit profile, outcome profile)?	Medium: (e.g., administration method, dosing schedule, benefit risk profile, outcome profile) Moderately complex, may require more tools to explain
Medical Context	How familiar is the patient with the medical setting (diagnostics involved, treatment in hospital, revalidation,...)?	Familiar/Simple Setting
Task	What type of task is being used? How much information is needed?	DCE consisting of multiple attributes with varying levels, some attributes are complex and require explanation

Table A2. Script of educational material and Attribute explanations as presented to patients

Diabetes and Glucose Monitoring Survey As someone with diabetes you probably already know a lot about blood glucose. Blood glucose is a sugar that the bloodstream carries to all cells in the body to supply energy. We get glucose into our bodies through eating and drinking. In people with diabetes, the body is unable to keep blood glucose at healthy levels. Because this can cause complications it is important that you check your blood glucose levels with a glucose monitor. Glucose monitors come in all different forms and functions. Today we would like to ask you questions about your preferences for glucose monitoring devices. This information will help to decide which devices should be available to patients and should be reimbursed by insurance companies.
We'd like to first give you some information about glucose monitoring. Glucose monitors are used to measure your blood glucose levels. In people without diabetes, blood glucose levels tend to stay relatively stable. This is because the body can make insulin itself and use insulin properly, so the amount of glucose in the blood is well regulated. In people with diabetes, the body is unable to keep blood glucose at healthy levels because it makes too little insulin or because the body does not use insulin properly. When this happens the blood glucose levels can become too high or too low. This can cause complications.
When your blood glucose is too low it is called hypoglycemia (or hypo for short). When you have hypoglycemia, you may start to feel unwell, your heart may race, you may sweat and your skin may turn pale. If your blood glucose is really low then you may have headaches or fall unconscious. When your blood glucose is too high it is called hyperglycemia (or hyper for short). When you have hyperglycemia, it can lead to trouble concentrating, slow reaction times, you may have a headache, and it may be difficult to focus your eyes. If your blood glucose is too high or too low for many years it can lead to serious complications in the long term. These complications could result in vision impairment or blindness, kidney disease, nerve damage, and heart disease. In some serious cases, these complications can lead to kidney dialysis, amputation of the foot, or heart attack.
To avoid complications, it is important to keep your blood glucose levels in a healthy range through healthy eating, physical activity, and medications. It is not always possible to tell if your blood glucose is at healthy levels based only on how you feel. Because of this, you can use a glucose monitor to check blood glucose levels yourself.
Today we want to ask you about specific aspects of glucose monitors to find out what is important to you. Glucose monitors come in all different forms and functions. You are probably most familiar with a finger-prick test. You prick your finger with a needle and lancing device, and place a drop of blood on a test strip in the monitor, and it will display your glucose level on the digital display. In addition, there are glucose monitors available that use sensors on your skin. These sensors continuously measure your glucose levels. There are two different types of sensors. First, there are continuous glucose monitors that automatically send your glucose levels to a small device or your phone or smartwatch. Second, there are flash glucose monitors that you need to swipe with a small device to get your glucose levels.
Today we want to ask you about specific aspects of glucose monitors to find out what is important to you. We will describe glucose monitors based on seven different aspects. **Precision** Some glucose monitors are more precise than others. Finger-pricking is generally regarded as the most accurate way to measure blood glucose levels. Measurements from devices that use sensors can be just as accurate as finger-pricking, but could also be less accurate. In our survey, the accuracy differs between different monitors and can have the following levels: As precise as finger prick Off by 6, meaning if your glucose level is 108 milligram per deciliter and you measure it with a device that is off by 6, then this device can say your glucose is anywhere from 102 to 114 milligram per deciliter. Off by 11, meaning: if your glucose level is 108 milligram per deciliter and you measure it with a device that is off by 11, then this device can say your glucose is anywhere from 97 to 119 milligram per deciliter.

Average number of finger-pricks per day

This is how many times you would need to do a finger-prick test on an average day. This is not the maximum number of times you can check your blood glucose levels you can check each day, but the recommended number of times to check. The average number of finger-pricks per day differs between different monitors and can have the following levels:

0 times a day

2 times a day

4 times a day

Effort to check

This means how much effort you need to give to check your blood glucose levels. The effort to check differs between different monitors and can have the following levels:

High effort checking means you need to stop what you're doing and concentrate on measuring your levels. You need to wash your hands, take out your device equipment, prick your finger, put blood on a strip, check the results, and then clean everything up.

Moderate effort checking means you need to keep a small dedicated device with you and get it out to scan the sensor on your body to obtain your glucose levels.

Low effort checking means your glucose levels are automatically sent to a device which you can view at any time. This could be a dedicated glucose device, your phone, or a smartwatch.

Chance of getting skin irritation or redness

Some sensors may cause skin irritation or redness around the sensor. This skin irritation or redness around a sensor can be uncomfortable and irritating and is similar to having an itchy allergic reaction. The sensor would likely need to be removed and replaced in a different spot or you may need to stop using it entirely. This skin irritation and redness usually lasts until after the sensor is replaced.

The chance of getting skin irritation or redness can be:

5% chance, meaning 5 out of a 100 people who get this device experience skin irritation and redness while 95 out of a 100 people do not experience this

20% chance, meaning 20 out of a 100 people who get this device experience skin irritation and redness while 80 out of a 100 people do not experience this

35% chance, meaning 35 out of a 100 people who get this device experience skin irritation and redness while 65 out of a 100 people do not experience this

Glucose Information

The glucose information given to you can be different between devices. When we say 'glucose information' we mean how your glucose levels are presented to you.

This information could be:

Your glucose level

Your glucose level with an arrow showing how your blood glucose is changing compared to your previous measurement. For example, increasing, decreasing, or staying the same.

Your current glucose level with a graphic of your blood glucose levels over the day.

Alarms

Some monitors can have an alarm (like a phone notification) any time your blood glucose levels are getting too high or too low. Devices can either have:

An alarm or

No alarm

Monthly costs

Monthly costs are how much money you need to pay out-of-pocket per month in order to check your blood glucose. This is money that is not reimbursed by your health insurance. This could be money needed to pay for devices, sensors, or strips used. These monthly costs could be:

€25 / 55 zł

€100 / 220 zł

€175 / 390 zł or

€250 / 550 zł

Chapter 9

Research Priorities to Increase Confidence in and Acceptance of Health Preference Research: What Questions Should be Prioritized Now?

Authors: Rachael L. DiSantostefano, Ian P. Smith, Marie Falahee, Aura Cecilia Jimenez-Moreno, Serena Oliveri, Jorien Veldwijk, G. Ardine de Wit, Ellen M. Janssen, Conny Berlin, Catharina GM Groothuis-Oudshoorn

Status: Published

DiSantostefano, R.L., Smith, I.P., Falahee, M., Jiménez-Moreno, A.C., Oliveri, S., Veldwijk, J., de Wit, G.A., Janssen, E.M., Berlin, C. and Groothuis-Oudshoorn, C.G., 2023. Research Priorities to Increase Confidence in and Acceptance of Health Preference Research: What Questions Should be Prioritized Now?. The Patient-Patient-Centered Outcomes Research, pp.1-12.

ABSTRACT

Background: There has been an increase in the study and use of stated-preference methods to inform medicines development decisions. The objective of this study was to create a prioritized research agenda of methodological questions relating to patient preferences based on the perspective of members of the preference research community.

Methods: Preference research stakeholders from industry, academia, consultancy, HTA/Regulatory, and patient organizations were surveyed about their perspectives on 19 methodological topics and questions for future studies that would increase acceptance of preference methods and their results by decision makers. The online survey consisted of an initial prioritization task, a best-worst scaling case 1 (BWS) instrument, and open-ended questions. The BWS used a balanced incomplete block design.

Results: 101 participants responded to the survey invitation with 66 completing the BWS responses and 2 partial BWS responses. The most important research topics related to a mix of applied and methodological research topics including synthesis of preferences across studies, transferability across populations or related diseases, and methods topics including comparison of methods and non-discrete choice experiment methods. Prioritization differences were found between respondents whose primary affiliation was academia vs. other stakeholders. Academic researchers prioritized methodological and/or less studied topics, whereas other stakeholders prioritized applied research topics relating to consistency of practice.

Conclusion: This study identified prioritized research topics that may help to increase confidence in both the robustness of preference methods and preference study results when applied to decision making across the medicine development lifecycle.

1. INTRODUCTION

The value of the patient perspective in the medical product lifecycle has never been more appreciated than it is at the current moment. Patients and patient advocacy groups, regulatory and Heath Technology Assessment (HTA) bodies, and industry leaders are increasingly advocating for the use of information collected from patients to inform product and trial designs, market access, and reimbursement schedules [1-4]. The Food and Drug Administration (FDA) has approved guidance on the use of patient preferences in marketing authorization [5]. The EMA gave a favorable opinion to a framework on planning and conducting patient preference studies [6], NICE has published a perspective on the use of preference data in HTA decision making [1], and the CIOMS working group XI published a report emphasizing the importance of including patient perspectives in medical product decision-making [7]. This has resulted in an ever-growing field of researchers who study patient preferences and an enormous growth in studies assessing what patients value in their healthcare [8]. With this interest in patient preference assessment there have also been calls to ensure that the studies are methodologically sound and produce reliable and valid information [9].

In order to address these issues, the IMI-PREFER project, a 6-year European public-private partnership, was launched in 2016 to inform on the use of patient preference studies for decision-making throughout the medical product lifecycle [10]. In 2018, the IMI-PREFER project conducted a survey to identify research priorities based on expert consensus, early literature reviews, stakeholder interviews, and a ranking exercise of research topics and questions [11]. The most important research priorities identified were related to four high-level concepts: evidentiary standards, assessment of preference heterogeneity, means to minimize patient burden, and means to maximize patient understanding of concepts presented in preference studies. These were used to guide the research questions addressed in 10 PREFER case studies which provided evidence to support recommendations on when and how to execute patient preference studies [12-22].

The field of preference research has evolved since these earlier prioritizations were completed with additional methods research being conducted over time [23, 24]. Thus, the objective of this study was to assess the updated research priorities of the preference community using an empirical approach.

2. METHODS

2.1 Participants and recruitment

Preference research stakeholders from industry, academia, consultancy, HTA/Regulatory, and four patient organizations were invited to participate in a one-time online survey. Invitations were sent through e-mail distribution lists of major health preference research groups, including the PREFER consortium (N=134), PREFER External Advisors (N=87),

the International Society for Pharmacoeconomics and Outcomes Research- Health Preference Research Special Interest Group (N=260), and (4) International Academy of Health Preference Research (N=143). Participants were invited to participate in a web-based survey and agreed to provide their expert opinion. There was no renumeration for participation. This study was not submitted for ethics approval as it elicited expert opinion and was not considered human subjects research. Data were analyzed in aggregate, and there was no attempt to identify individuals based on individual characteristics provided.

Recruitment started April 1, 2022 and completed May 16. Potential respondents were sent an initial email followed by two reminders. The number of participants was not capped and a minimum sample of at least 50 participants was deemed sufficient to allow for exploration of heterogeneity based on prior research [25, 26].

2.2 Objects: Research Topic Identification

Objects used in the prioritization tasks were research topics that could increase the confidence in and acceptance of patient preference research in decision-making throughout the medical product lifecycle by organizations and groups like government regulators (EMA, FDA), reimbursement agencies, patient groups, and industry. These objects were identified in line with good-research practices [27], including a review of previous research agendas [11, 23, 24], consideration of important insights and publications from the literature, and solicitation from PREFER consortium members and scientific advisors.

Nineteen objects and corresponding descriptions were finalized in consideration of pre-testing (Table 1). Each object was given a short name followed by a more detailed description to ensure the research topic was understood uniformly by all participants. Objects were reviewed by the co-authors of the study for clarity and by nine preference researchers outside of the research team including five that were not involved in PREFER (see Acknowledgements).

2.3 Survey instrument

The survey began with background questions related to the participants' professional affiliation, familiarity with preference research, and geographic location. Respondents were then presented with the 19 methodological research topics and asked to complete two prioritization tasks. The first task consisted of classifying the 19 topics into four importance categories ('Important question to study in future'; 'Important but studied adequately already'; 'Important but too complicated or impossible to study'; 'Not important to study in future'). The second task was a Best-Worst Scaling (BWS-1) exercise in which participants were asked to select Best (most important) and Worst (least important) topics for future studies that would "increase acceptance of preference methods and their results by decision makers". Respondents were initially presented with an example Best-Worst Scaling 1 (BWS-1) choice task and then asked

to complete 19 BWS-1 tasks. The BWS-1 design used a pre-specified balanced incomplete block design [28], and each BWS-1 task presented four objects to the participant in random order. During the choice tasks, participants could view objects and their descriptions.

Table 1. Research topics and example questions assessed in exercises

1.	**Comparing methods**: How do the preference mean results and preference heterogeneity results of different patient preference methods compare when applied to address the same research question using the same attributes and samples from the same population?
2.	**Changing number of attributes**: How do changes in the number or types (e.g., categorical vs. numerical value) of attributes impact results for a given method?
3.	**Attribute presentation & framing**: How do changes in the framing (e.g., mortality vs. survival) and attribute presentation (e.g., graphical representation of risk versus text) impact results for a given method?
4.	**Transferability across populations or related diseases**: How transferable are preferences from one specific disease population to another population (e.g., related diseases, different diseases but similar complaints, same disease but different countries)?
5.	**Method selection guidance**]: How to determine which preference assessment method to use in a given context, patient population or for a specific research purpose?
6.	**Educational Materials – Which Material to Enhance?** What information (e.g., risk information, disease context) benefits most from the use of enhanced educational material (such as videos, voiceovers, gamification, and animations) to inform patients?
7.	**Educational Materials – Digital v. Text Formats**: How do different types of enhanced educational material (such as videos, voiceovers, gamification, and animations) affect engagement, understanding, choice consistency, and preferences compared to static text and images?
8.	**Educational Materials – Low Literacy and Numeracy** What types of educational materials are optimal for samples where low literacy and/or low numeracy may be prevalent?
9.	**Internal Validity / Data Quality**: How should one best assess whether patients understand and are paying attention to a given set of cognitive tasks?
10.	**Psychological Constructs – explain preferences/heterogeneity**: In which situations do psychosocial constructs (e.g., personal beliefs/personality traits or attitudes) have value in explaining preferences and preference heterogeneity?
11.	**Psychological Constructs – explain preferences across methods**: To what extent are relationships between measures of psychological constructs and patient preferences consistent across preference elicitation methodologies (e.g., are relationships between psychological constructs and preferences found with a DCE similar to the relationships found between psychological constructs with Best-Worst Scaling?)?
12.	**Changes in preferences over time**: Which factors influence the stability of preferences over time and why? (e.g., changes in health states, adjustment to condition, nature of illness and treatment, and changes in knowledge)?
13.	**Individual preferences**: How can individual preferences be used in shared decision-making, (e.g., in the development of decision aids or value clarification)?
14.	**Synthesis of Preferences Across Studies**: How to best synthesize multiple patient preference studies for either meta-analysis or predicting preferences for a particular context?
15.	**Mapping patient-reported outcomes to preference study attributes**: How can attributes in a patient preference study be mapped to patient-reported outcomes (or clinical outcome assessments in general)? (e.g., in mapping preferences to a patient-reported outcome in a clinical trial, or incorporating a patient-reported outcome within a preference study)?
16.	**Revealed preferences – role in decision making**: When and how might revealed preferences be used for decision making in the medical product lifecycle?
17.	**Revealed preferences – external validity**: How well do stated preferences match revealed preferences in different disease areas or health care decisions, and under what conditions would we expect them to differ?
18.	**Expressing uncertainty in patient preference studies**: When and how should uncertainty around benefit and risk estimates be incorporated into the design of patient preference studies?
19.	**Study non-DCE Methods**: Develop evidence-based good-research practices on the conduct, analyses and use of non-DCE preference methods (e.g., Best-Worst Scaling Types 1 – 3, Swing Weighting, Probabilistic Threshold Technique).

As medical products in preference tasks are often described using attributes which are applicable to multiple different treatments and disease areas, respondents were asked questions about the utility of an attribute library for reference in patient preference research following the prioritization tasks, ("Do you think an attribute library would be a useful contribution to the field?"). Finally, an open-text question was asked in which respondents could comment on research priorities and additional research topics. Respondents were able to navigate forward and backward within the survey.

The final survey was pre-tested online by co-authors to test the survey, remove software bugs, and make any final wording adjustments to the survey or instructions to improve clarity. Analysis of results was done in R [29].

2.4 Data analysis

Descriptive statistics were conducted on participant characteristics and prioritization tasks. The initial classification task about whether a research topic was important to study in the future was analyzed by comparing the proportions in each response category (not important, important, important but studied adequately already, etc.). The analysis of BWS-1 consisted of three different tabulations of choice frequencies for each research topic: the number of best selections, the number of worst selections, and the best-worst score calculated as the difference between number of best and number of worst selections which incorporates differences in opinion into topic prioritization. These analyses were also conducted to compare sub-group priorities between respondents who reported as being academically affiliated versus all other stakeholders. No formal significance tests were conducted to test for differences between groups.

3. RESULTS

3.1 Participant Characteristics

Of the N=107 participants responding to the survey, n=101 completed the demographics, n=76 completed the initial ranking exercise, and n=66 participants completed the BWS-1 choice task (Table 2). Response rates could not be calculated due to the overlap in membership of n=32 respondents asked to complete the survey. The n=33 respondents that dropped out before the BWS-1 included participants from all stakeholder groups including n=6 from industry, n=14 from academia, n=3 from consultancy, n=2 from HTA, and n=7 from patient organizations. Among those that completed the survey, median completion time was about 20 minutes.

Table 2. Respondent Demographics (%)

Demographic Characteristic	All Respondents n=101	%
Patient Preference Work Area		
Academia	45	45
Industry	27	27
Consultancy	10	10
Regulatory Agency	5	5
HTA	4	4
Patient Organizations	9	9
Other	0	0
Professional Community Membership (multiple selections allowed)		
PREFER	48	48
International Academy of Health Preference Research	38	38
ISPOR Health Preference Research SIG	37	37
BRACE SIG	8	8
Other	6	6
No answer selected	8	8
Multiple group memberships	32	32
PREFER Case Study involvement (multiple selections allowed)		
None	64	63
Core case study or studies (Lung cancer, RA, NMD)	26	26
Academic case study or studies	15	15
Industry case study or studies	8	8
No answer selected	2	2
Familiarity with patient preference studies (multiple responses allowed)		
I was not aware of what patient preferences studies were before this survey	0	0
I have read about patient preference studies (e.g., manuscript, report, protocol)	62	61
I have peer reviewed patient preference studies	57	56
I have attended webinars/conference session on patient preference studies	76	75
I have organized, designed, or managed patient preference studies	64	63
I have performed analyses of patient preference study data	55	54
I have used the results of patient preference studies in my work	55	54
I have other experience with preference studies	6	6
World Bank Region		
East Asia and Pacific	7	7
Europe and Central Asia	52	51
Latin America & the Caribbean	1	1
Middle East and North Africa	2	2
North America	37	37
South Asia	1	1
Sub-Saharan Africa	0	0

HTA: Health Technology Assessment; SIG: Special Interest Group; Under the category 'Other' the following communities where mentioned: MDIC, iHEA/iHEA SIG, PFMD, National Community; 'Other' the following was mentioned: I have conducted a wide range of methods work around preference elicitation, I worked on training for PREFER, commissioning of patient preference studies, I have evaluated studies from a regulatory science perspective, health valuation, non-patient preferences, Public health and patient access.Note: Of the 66 respondents who completed the BWS, the patient preference work areas reported were Academia (n=29), Industry (n=21), Consultancy (n=7), Regulatory Agency (n=5), HTA (n=2) or patient organizations (n=2), and 68% reported having organized, designed, or managed patient preference studies, 59% reported having performed analyses of patient preference study data, and 67% reported having used the results of patient preference studies in their work.

3.2 Direct Categorization of Research Topics

The results of the first categorization task are presented in in Figure 1. In all cases, the topics presented were identified as being important for future studies by a majority of respondents. The topics categorized the most as 'important questions to study in future' were *Synthesis of Preferences Across Studies* and *Expressing uncertainty in patient preference studies*. The topics that were most categorized as 'important but studied adequately already' were related to *Internal Validity / Data Quality*, *Attribute presentation & framing*, and *Method selection guidance*. Three topics (*Transferability across populations or related diseases*, *Changes in preferences over time*, and using *Psychological Constructs to explain preferences across methods*) received the most categorizations of being 'important but too complicated or impossible to study'. Using *Psychological Constructs to explain preferences across methods* also received the most categorizations of 'not important to study in future' along with *Revealed preferences - role in decision making*, and the impact of *Changing number of attributes*.

3.3 BWS-1 Ranking of Research Topics

The results of the BWS-1 are presented in figure 2 and supplementary materials. The topics which were selected as 'Best' the most were *Transferability across populations or related diseases*, *Comparing methods*, *Synthesis of Preferences Across Studies*, and *Method selection guidance*. The topics which received the most 'Worst' selections were *Revealed preferences - role in decision making*, *Psychological Constructs - explain preferences across methods*, *Educational Materials - Digital v. Text Formats*, and *Changing number of attributes*. The topics which had the highest ratio of Best-Worst selections were *Method selection guidance*, *Changes in preferences over time*, *Synthesis of Preferences Across Studies*, *Transferability across populations or related diseases,* and *Internal Validity / Data Quality*.

3.4 Comparison of respondents with an academic affiliation versus other stakeholders

Different priorities were found in the two tasks between respondents with primary academic affiliation vs. other affiliations. In the categorization exercise, other stakeholders were more likely to label a topic or question as too difficult or impossible to research compared to those working inside academia (Figure 3).

In the BWS-1, respondents working within academia tended to prioritize more methodological or less studied topics, such as Transferability, and External validity, while those working outside academia tended to prioritize more applied research topics relating to improving? consistency of practice, such as Methods selection guidance, Internal validity, and Synthesis of preferences across studies (Figure 3). No difference was found in the lowest five priority topics between those working in academia and other stakeholder groups (*Revealed preferences- role in decision making, Educational materials – which materials to enhance, Education materials – digital v. text formats, and Psychological constructs – explain preferences across methods, in the lowest five priority topics*).

Figure 1. Categorizations for research topics and questions (n=76)

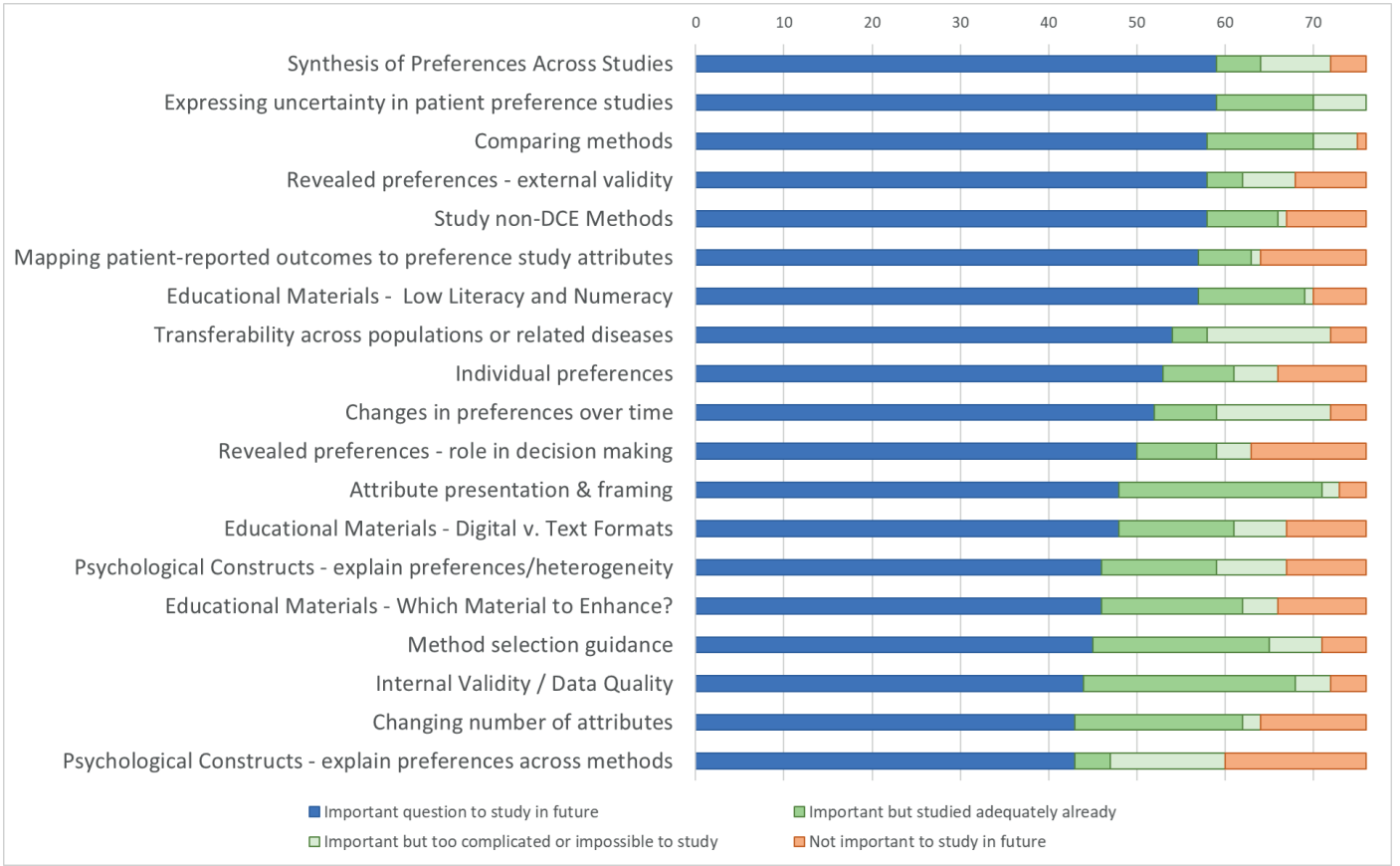
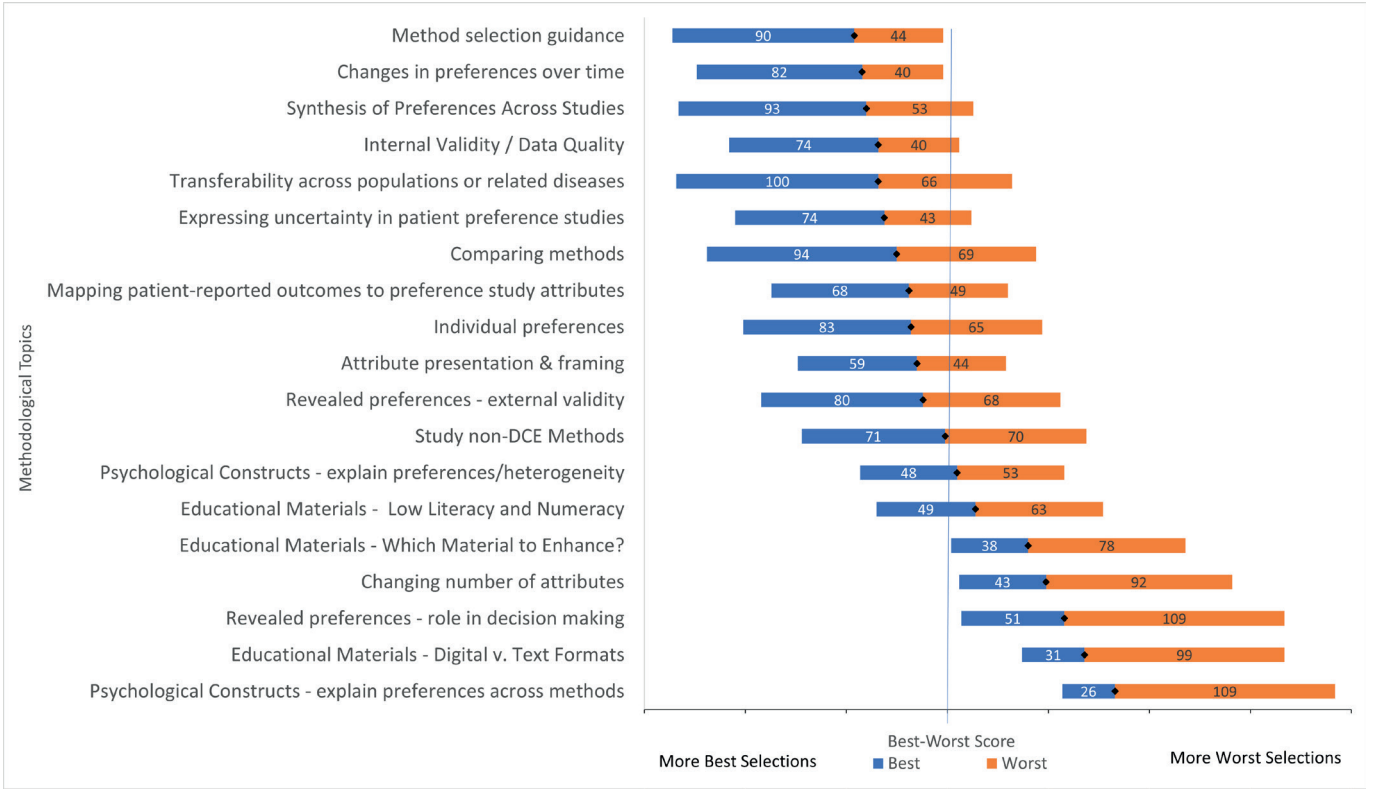


Figure 2. Best-Worst Score for each Research Topic (n=66)



Note: Number of times selected as 'Best' topic (blue) or 'Worst' topic (orange) shown in bar. Topics are sorted from highest Best-Worst difference to lowest Best-Worst difference. Light blue line indicates an even number of 'Best' and 'Worst' selections.

Figure 3. BWS-1 Ranking by Academia (left, n=37) vs. Other Stakeholders (right, n=29)

3.4 Preference Attribute Libraries

Sixty-six (62%) respondents responded to the question regarding the usefulness of an attribute library. Of those, n=52 (79%) responded 'yes', n=4 (6%) responded 'no', and n=10 (15%) responded 'not sure'. N=30 (48%) considered the attribute library as equally important or more important than the 19 research topics and questions presented in the prioritization tasks. The therapeutic areas most often given where an attribute library would be beneficial were oncology (n=15), rare diseases and cardiovascular diseases (n=4 each), and vaccination or infectious diseases and chronic diseases (n=2 each). Respondents found the possibility of an attribute library important for specific benefits and risks related to commonly used endpoints or those that have been established as "gold-standard" endpoints or concepts (e.g., mortality, survival and progression free survival in oncology). However, respondents raised concerns about attributes being context dependent and not necessarily re-usable, or raised concerns about feasibility (e.g., how to standardize attributes and maintain the library). Two respondents referred to existing frameworks that could cover or contribute to the creation of an attribute library (a disease-specific Core Outcome Set and EuroQoL).

3.5 Free Text Comments about Research Priorities and the Survey

N=20 respondents answered the open-ended questions "Is there anything else you want to share about future research priorities in patient preference research to increase acceptance of these methods?". N=7 participants commented positively and found research topics included in the ranking exercise comprehensive. Respondents reported the following topics as more or equally important as the topics in the list: Account for preference heterogeneity using patient's personal aspects, context, and other social determinants of health (5 mentions); Neutral entities to perform PP studies – which avoids potential biases and may ensure methodological rigor (2 mentions); Use preferences to guide endpoint selection in clinical trials (1 mention); Include under-represented populations in PP studies (1 mention). Additional remarks raised by respondents can be summarized into 2 main areas: establishing responsible entities for performing PP studies that give confidence in robustness/validity of methods (mentioned 3 times); need for guidelines and best practice in PP studies (mentioned 5 times).

4. DISCUSSION

Over the past decades, there has been increased interest in measuring patient preferences to aid decision making during drug development. This has generated questions about how to assess patient preferences reliably from a variety of different stakeholders (including academic researchers, industry members, consultancies, health authorities, and patient groups). This prioritization exercise was conducted to prioritize research topics for the health preference research community with the goal of increasing acceptance of patient preference methods and their results by decision makers in the medical product lifecycle. We identified 19 important

research topics for future study that would increase acceptance of preference methods and their results by decision makers. Within these 19 topics there were clear priorities for specific topics.

While all the research topics presented were considered important to study by a majority of the respondents, the most important research topics related to a mix of methodological and applied research topics. Two of the highest priority topics were related to the use of patient preference research outside of the individual study population: either for use in other populations or for use in meta-analysis and predicting preferences. Both of these topics were identified as important in both ranking tasks and were not listed as having been previously studied. Conducting a preference study can be a time and resource intensive undertaking, so the reuse of previous patient preferences to inform new or future decision-making can help ensure that patient values are considered when a new study is not possible or necessary.

Five topics (*Internal Validity / Data Quality*, *Attribute presentation & framing*, *Method selection guidance*, *Changing number of attributes*, and *Educational Materials - Which Material to Enhance?*) were listed as important for future research by over half of the respondents, though over 20% of respondents felt that these topics had been researched enough previously. The topic of *Method selection guidance* was highly rated in the BWS-1 task indicating that many respondents think it is a top priority despite 34% of respondents thinking it had already been researched adequately. The disparity between ranking these as important topics and thinking that while important it has previously been studied enough may simply be a difference of opinion. However, it may also reflect a lack of awareness of previous work in this area. Recent publications have highlighted decision criteria that can be used to guide method selection [30, 31], and previous research has been published on internal validity tests and patient comprehension [32-34], attribute presentation and framing [35-38], the number of attributes [39], and educational materials [16, 40, 41] so this finding may reflect variability in awareness of this previous work to help. As the amount of methodological research available increases, there will be a need to provide consolidated and updated dissemination resources. Examples of these types of resources could be online courses and webinars, seminars, trainings hosted by professional organizations, or catalogues and repositories of published studies.

Survey participants did express interest in a library of previously developed attributes for targeted areas, for example, oncology outcomes and outcomes frequently seen across diseases. Challenges with an attribute library include reaching consensus on which attributes to include and the most appropriate attribute definitions, sufficient uptake, and long-term sustainability. One possible model towards an attribute library could be to follow the example of OMERACT[42], an independent organization that strives to improve endpoint outcomes through a data-driven, iterative consensus process involving relevant stakeholder groups. In addition, an intermediate step towards an attribute library could be the registration of most preference studies in a standardized way, for example through the Health Preference Study Technology Registry (HPSTR)[43].

If we compare this study's results to the previous PREFER prioritization exercise, some trends can be observed [11]. In the previous prioritization exercise, transferability of preference results both within a patient population and to other populations were highly prioritized. These topics overlap with the current study topics of synthesis of preferences across studies and transferability across populations or related diseases, which were more highly prioritized than in the original study indicating that they remain a topic of interest to stakeholders. Additionally, the comparison of different methods for preference elicitation, study of non-DCE methods, and consistency of preference outcomes from different methods were highly prioritized topic in the previous exercise. In the current study this topic remained an important topic with indications that while additional research has been done in these areas since the first prioritization exercise [30, 44-47], more research would help to understand which method to choose when conducting a preference study. Interestingly, the topic of stability of preferences over time was previously ranked as least important, but in this updated exercise it was considered the second most important research topic.

Not unexpectedly, there were differences in prioritization of patient preference research based on stakeholder affiliation. Academic researchers tended to prioritize methodological and/or less studied topics, including transferability, and external validity. Other stakeholders, most of them are likely to use preferences for decision-making, prioritized applied research topics relating to consistency of practice, including methods selection guidance, internal validity, and synthesis of preferences across studies. Differences in prioritization of research topics most likely reflect the different needs of different stakeholders. To ensure that the needs of a variety of stakeholders are met and to encourage a diversity of perspectives, it remains important to continue cross-sector collaborations.

Lower priority topics were similar across the groups, including questions on revealed preferences, educational materials, and psychological constructs. Lower prioritization of these topics may reflect the perception that these questions are difficult, if not impossible, to answer (Revealed preferences) or lower levels of familiarity with topics by members of the preference research community (Educational materials and Psychological constructs).

A strength of this study was the use of multiple instruments to rank the topics, allowing for a richer understanding of respondent opinions. By combining the results, we were able to understand why some topics may have been important to some while not being prioritized by others. Another strength of this study was the inclusion of the broader patient preference community, including professional society preference research groups and scientific advisors beyond PREFER researchers. This differs from previous prioritization exercises in that it included a broader community. However, this study did have some limitations. One limitation of this study was that the study topics were identified based on experiences with PREFER by researchers involved with PREFER. In line with this, the survey was sent to professional societies and mailing lists that had existing professional relationships with the study authors.

The survey was not publicized outside of these networks and did not attempt to recruit preference researchers not affiliated with these professional societies. The sample therefore reflects a convenience sample.

There is great promise in the use of PPS to inform decisions across the medical product lifecycle, and future research topics should be prioritized to bolster confidence in and use of these methods. Our study proposes an agenda for future research. We encourage preference researchers to continue contributing toward research needs as prioritized with this study and to increasing the confidence in both the robustness of preference methods and preference study results when applied to decision making across the medicine development lifecycle.

REFERENCES

1. Bouvy, J.C., et al., *Use of Patient Preference Studies in HTA Decision Making: A NICE Perspective.* The Patient - Patient-Centered Outcomes Research, 2020. **13**(2): p. 145-149.
2. Mott, D.J., *Incorporating Quantitative Patient Preference Data into Healthcare Decision Making Processes: Is HTA Falling Behind?* The Patient - Patient-Centered Outcomes Research, 2018. **11**(3): p. 249-252.
3. Ho, M., et al., *A Framework for Incorporating Patient Preferences Regarding Benefits and Risks into Regulatory Assessment of Medical Technologies.* Value in Health, 2016. **19**(6): p. 746-750.
4. Haerry, D., et al., *EUPATI and Patients in Medicines Research and Development: Guidance for Patient Involvement in Regulatory Processes.* Frontiers in Medicine, 2018. **5**.
5. US Food Drug Administration, *Patient preference information—voluntary submission, review in premarket approval applications, humanitarian device exemption applications, and de novo requests, and inclusion in decision summaries and device labeling*, in *Food and Drug Administration staff, and other stakeholders*. 2016.
6. Committee for Medicinal Products for Human Use (CHMP), *Qualification Opinion of IMI PREFER*. 2022.
7. CIOMS Working Group report, *Patient involvement in the development, regulation and safe use of medicines*. 2022: Geneva, Switzerland.
8. Soekhai, V., et al., *Discrete Choice Experiments in Health Economics: Past, Present and Future.* Pharmacoeconomics, 2019. **37**(2): p. 201-226.
9. Janssens, R., et al., *Opportunities and challenges for the inclusion of patient preferences in the medical product life cycle: a systematic review.* BMC Med Inform Decis Mak, 2019. **19**(1): p. 189.
10. de Bekker-Grob, E.W., et al., *Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public-Private Project.* Patient, 2017. **10**(3): p. 263-266.
11. Smith, I.P., et al., *Methodological Priorities for Patient Preferences Research: Stakeholder Input to the PREFER Public-Private Project.* Patient, 2021. **14**(5): p. 449-453.
12. The PREFER Consortium, *PREFER Recommendations - Why, when and how to assess and use patient preferences in medical product decision-making*. 2022.
13. Simons, G., et al., *Preferences for preventive treatments for rheumatoid arthritis: discrete choice survey in the UK, Germany and Romania.* Rheumatology, 2022.
14. Jimenez-Moreno, A.C., et al., *Patient Preferences in Rare Diseases: A Qualitative Study in Neuromuscular Disorders to Inform a Quantitative Preference Study.* The Patient - Patient-Centered Outcomes Research, 2021. **14**(5): p. 601-612.
15. Monzani, D., et al., *Patient Preferences for Lung Cancer Treatments: A Study Protocol for a Preference Survey Using Discrete Choice Experiment and Swing Weighting.* Frontiers in Medicine, 2021. **8**.
16. Bywall, K.S., et al., *Does being exposed to an educational tool influence patient preferences? The influence of an educational tool on patient preferences assessed by a discrete choice experiment.* Patient Education and Counseling, 2021. **104**(10): p. 2577-2585.
17. Janssens, R., et al., *Patient Preferences for Multiple Myeloma Treatments: A Multinational Qualitative Study.* Frontiers in Medicine, 2021. **8**.
18. Cook, N.S., et al., *People living with moderate-to-severe COPD prefer improvement of daily symptoms over the improvement of exacerbations: a multicountry patient preference study.* ERJ Open Research, 2022. **8**(2): p. 00686-2021.

19. Turk, D., et al., *Patient preferences for osteoarthritis pain and chronic low back pain treatments in the United States: a discrete-choice experiment.* Osteoarthritis and Cartilage, 2020. **28**(9): p. 1202-1213.

20. Pinto, C.A., et al., *Comparing Patient Preferences for Antithrombotic Treatment During the Acute and Chronic Phases of Myocardial Infarction: A Discrete-Choice Experiment.* The Patient - Patient-Centered Outcomes Research, 2022. **15**(2): p. 255-266.

21. van Overbeeke, E., et al., *Patient preferences for gene therapy in haemophilia: Results from the PAVING threshold technique survey.* Haemophilia, 2021. **27**(6): p. 957-966.

22. Smith, I.P., et al., *Diabetes patient preferences for glucose-monitoring technologies: results from a discrete choice experiment in Poland and the Netherlands.* BMJ Open Diabetes Res Care, 2023. **11**(1).

23. Levitan, B., et al., *The Ball is in Your Court: Agenda for Research to Advance the Science of Patient Preferences in the Regulatory Review of Medical Devices in the United States.* Patient, 2017. **10**(5): p. 531-536.

24. Meara, A., N.L. Crossnohere, and J.F.P. Bridges, *Methods for measuring patient preferences: an update and future directions.* Curr Opin Rheumatol, 2019. **31**(2): p. 125-131.

25. Janssen, E.M., et al., *Identifying and prioritizing concerns associated with prosthetic devices for use in a benefit-risk assessment: a mixed-methods approach.* Expert Rev Med Devices, 2018. **15**(5): p. 385-398.

26. Cheung, K.L., et al., *Using Best-Worst Scaling to Investigate Preferences in Health Care.* Pharmacoeconomics, 2016. **34**(12): p. 1195-1209.

27. Bridges, J.F., et al., *Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force.* Value Health., 2011. **14**(4): p. 403-13. doi: 10.1016/j.jval.2010.11.013. Epub 2011 Apr 22.

28. Marley, A.A.J., J.J. Louviere, and T.N. Flynn, *The BWS object case*, in *Best-Worst Scaling: Theory, Methods and Applications*. 2015, Cambridge University Press: Cambridge. p. 14-55.

29. ChoiceMetrics, *Ngene 1.1.1*. 2012, User Manual & Reference Guide, Australia.

30. Soekhai, V., et al., *Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review.* Drug Discovery Today, 2019. **24**(7): p. 1324-1331.

31. Whichello, C., et al., *Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison.* BMC Medical Informatics and Decision Making, 2020. **20**(1): p. 114.

32. Johnson, F.R., J.-C. Yang, and S.D. Reed, *The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments.* Value in Health, 2019. **22**(2): p. 157-160.

33. Janssen, E.M., et al., *Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?* Expert Review of Pharmacoeconomics & Outcomes Research, 2017. **17**(6): p. 531-542.

34. Pearce, A., et al., *Respondent Understanding in Discrete Choice Experiments: A Scoping Review.* Patient, 2021. **14**(1): p. 17-53.

35. Smith, I.P., et al., *Does It Matter How You Ask? Assessing the Impact of Failure or Effectiveness Framing on Preferences for Antibiotic Treatments in a Discrete Choice Experiment.* Patient Prefer Adherence, 2022. **16**: p. 2921-2936.

36. Veldwijk, J., et al., *Survival or Mortality: Does Risk Attribute Framing Influence Decision-Making Behavior in a Discrete Choice Experiment?* Value Health, 2016. **19**(2): p. 202-9.

37. Edwards, A., et al., *Presenting risk information--a review of the effects of "framing" and other manipulations on patient outcomes.* J Health Commun, 2001. **6**(1): p. 61-82.

38. Vass, C., D. Rigby, and K. Payne, *"I Was Trying to Do the Maths": Exploring the Impact of Risk Communication in Discrete Choice Experiments.* The Patient - Patient-Centered Outcomes Research, 2019. **12**(1): p. 113-123.

39. Spinks, J. and D. Mortimer, *Lost in the crowd? Using eye-tracking to investigate the effect of complexity on attribute non-attendance in discrete choice experiments.* BMC Medical Informatics and Decision Making, 2016. **16**(1): p. 14.

40. Lim, S.L., et al., *Are Videos or Text Better for Describing Attributes in Stated-Preference Surveys?* The Patient - Patient-Centered Outcomes Research, 2020. **13**(4): p. 401-408.

41. Vass, C.M., et al., *A picture is worth a thousand words: the role of survey training materials in stated-preference studies.* The Patient-Patient-Centered Outcomes Research, 2020. **13**(2): p. 163-173.

42. Tugwell, P., et al., *OMERACT: an international initiative to improve outcome measurement in rheumatology.* Trials, 2007. **8**: p. 38.

43. International Academy of Health Preference Research. *Health Preference Study and Technology Registry*. Available from: https://hpstr.org/landing.

44. Veldwijk, J., et al., *Suitability of Preference Methods Across the Medical Product Lifecycle: A Multicriteria Decision Analysis.* Value Health, 2022.

45. Whichello, C., et al., *An overview of critical decision-points in the medical product lifecycle: Where to include patient preference information in the decision-making process?* Health Policy, 2020. **124**(12): p. 1325-1332.

46. van Dijk, J.D., et al., *An Empirical Comparison of Discrete Choice Experiment and Best-Worst Scaling to Estimate Stakeholders' Risk Tolerance for Hip Replacement Surgery.* Value in Health, 2016. **19**(4): p. 316-322.

47. Himmler, S., et al., *What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting.* Journal of Choice Modelling, 2021. **38**: p. 100265.

APPENDIX

Appendix Table A1. BWS/Importance Results ranked according to highest Best-Worst ratio (loaded separately)

item	Levels	Long	Important	Done already	Too difficult	Not important	B-W	Best	Worst	times	sum	rank
5	Method selection guidance	How to determine which preference assessment method to use in a given context, patient population or for a specific research purpose?	26	9	3	4	40	62	22	148	1000	10
9	Internal Validity / Data Quality	How should one best assess whether patients understand and are paying attention to a given set of cognitive tasks?	22	14	3	3	29	48	19	148	725	20
14	Synthesis of Preferences Across Studies	How to best synthesize multiple patient preference studies for either meta-analysis or predicting preferences for a particular context?	31	4	5	2	26	51	25	148	650	30
3	Attribute presentation & framing	How do changes in the framing (e.g., mortality vs. survival) and attribute presentation (e.g., graphical representation of risk versus text) impact results for a given method?	26	12	2	2	20	37	17	148	500	45
15	Mapping patient-reported outcomes to preference study attributes	How can attributes in a patient preference study be mapped to patient-reported outcomes (or clinical outcome assessments in general)? (e.g., in mapping preferences to a patient-reported outcome in a clinical trial, or incorporating a patient-reported outcome within a preference study)?	33	3	1	5	20	44	24	148	500	45
12	Changes in preferences over time	Which factors influence the stability of preferences over time and why? (e.g., changes in health states, adjustment to condition, nature of illness and treatment, and changes in knowledge)	25	4	10	3	19	44	25	148	475	60
1	Comparing methods	How do the preference mean results and preference heterogeneity results of different patient preference methods compare when applied to address the same research question using the same attributes and samples from the same population?	32	4	5	1	17	57	40	148	425	70

item	Levels	Long	Important	Done already	Too difficult	Not important	B-W	Best	Worst	times	sum	rank
19	Study non-DCE Methods	Develop evidence-based good-research practices on the conduct, analyses and use of non-DCE preference methods (e.g., Best-Worst Scaling Types 1 - 3, Swing Weighting, Probabilistic Threshold Technique).	34	5	1	2	13	48	35	148	325	80
18	Expressing uncertainty in patient preference studies	When and how should uncertainty around benefit and risk estimates be incorporated into the design of patient preference studies?	30	6	6	0	11	40	29	148	275	90
4	Transferability across populations or related diseases	How transferable are preferences from one specific disease population to another population (e.g., related diseases, different diseases but similar complaints, same disease but different countries)	27	3	10	2	-2	43	45	148	-50	100
8	Educational Materials - Low Literacy and Numeracy	What types of educational materials are optimal for samples where low literacy and/or low numeracy may be prevalent?	29	7	1	5	-3	28	31	148	-75	110
2	Changing number of attributes	How do changes in the number or types (e.g., categorical vs. numerical value) of attributes impact results for a given method?	27	6	2	7	-5	35	40	148	-125	120
10	Psychological Constructs - explain preferences/ heterogeneity	In which situations do psychosocial constructs (e.g., personal beliefs/personality traits or attitudes) have value in explaining preferences and preference heterogeneity?	20	9	6	7	-10	21	31	148	-250	130
13	Individual preferences	How can individual preferences be used in shared decision-making, (e.g., in the development of decision aids or value clarification)?	26	6	3	7	-12	36	48	148	-300	140
17	Revealed preferences - external validity	How well do stated preferences match revealed preferences in different disease areas or health care decisions, and under what conditions would we expect them to differ?	28	2	5	7	-18	30	48	148	-450	150

item	Levels	Long	Important	Done already	Too difficult	Not important	B-W	Best	Worst	times	sum	rank
6	Educational Materials - Which Material to Enhance?	What information (e.g., risk information, disease context) benefits most from the use of enhanced educational material (such as videos, voiceovers, gamification, and animations) to inform patients?	23	9	3	7	-23	21	44	148	-575	160
7	Educational Materials - Digital v. Text Formats	How do different types of enhanced educational material (such as videos, voiceovers, gamification, and animations) affect engagement, understanding, choice consistency, and preferences compared to static text and images?	24	7	5	6	-38	18	56	148	-950	170
11	Psychological Constructs - explain preferences across methods	To what extent are relationships between measures of psychological constructs and patient preferences consistent across preference elicitation methodologies (e.g., are relationships between psychological constructs and preferences found with a DCE similar to the relationships found between psychological constructs with Best-Worst Scaling?)?	20	2	10	10	-40	14	54	148	-1000	180
16	Revealed preferences - role in decision making	When and how might revealed preferences be used for decision making in the medical product lifecycle?	26	5	3	8	-44	26	70	148	-1100	190

Appendix Table A2. Academic BWS/Importance Results ranked according to highest Best-Worst ratio

item	Levels	Long	Important	Done already	Too difficult	Not important	B/W	Best	Worst	times	sum	rank
4	Transferability across populations or related diseases	How transferable are preferences from one specific disease population to another population (e.g., related diseases, different diseases but similar complaints, same disease but different countries)	27	1	4	2	36	57	21	116	900	10
13	Individual preferences	How can individual preferences be used in shared decision-making, (e.g., in the development of decision aids or value clarification)?	27	2	2	3	30	47	17	116	750	25
17	Revealed preferences - external validity	How well do stated preferences match revealed preferences in different disease areas or health care decisions, and under what conditions would we expect them to differ?	30	2	1	1	30	50	20	116	750	25
12	Changes in preferences over time	Which factors influence the stability of preferences over time and why? (e.g., changes in health states, adjustment to condition, nature of illness and treatment, and changes in knowledge)	27	3	3	1	23	38	15	116	575	40
18	Expressing uncertainty in patient preference studies	When and how should uncertainty around benefit and risk estimates be incorporated into the design of patient preference studies?	29	5	0	0	20	34	14	116	500	50
14	Synthesis of Preferences Across Studies	How to best synthesize multiple patient preference studies for either meta-analysis or predicting preferences for a particular context?	28	1	3	2	14	42	28	116	350	60
1	Comparing methods	How do the preference mean results and preference heterogeneity results of different patient preference methods compare when applied to address the same research question using the same attributes and samples from the same population?	26	8	0	0	8	37	29	116	200	70

item	Levels	Long	Important	Done already	Too difficult	Not important	B/W	Best	Worst	times	sum	rank
5	Method selection guidance	How to determine which preference assessment method to use in a given context, patient population or for a specific research purpose?	19	11	3	1	6	28	22	116	150	80
9	Internal Validity / Data Quality	How should one best assess whether patients understand and are paying attention to a given set of cognitive tasks?	22	10	1	1	5	26	21	116	125	95
10	Psychological Constructs - explain preferences/ heterogeneity	In which situations do psychosocial constructs (e.g., personal beliefs/personality traits or attitudes) have value in explaining preferences and preference heterogeneity?	26	4	2	2	5	27	22	116	125	95
15	Mapping patient-reported outcomes to preference study attributes	How can attributes in a patient preference study be mapped to patient-reported outcomes (or clinical outcome assessments in general)? (e.g., in mapping preferences to a patient-reported outcome in a clinical trial, or incorporating a patient-reported outcome within a preference study)?	24	3	0	7	-1	24	25	116	-25	110
3	Attribute presentation & framing	How do changes in the framing (e.g., mortality vs. survival) and attribute presentation (e.g., graphical representation of risk versus text) impact results for a given method?	22	11	0	1	-5	22	27	116	-125	120
8	Educational Materials - Low Literacy and Numeracy	What types of educational materials are optimal for samples where low literacy and/or low numeracy may be prevalent?	28	5	0	1	-11	21	32	116	-275	130
19	Study non-DCE Methods	Develop evidence-based good-research practices on the conduct, analyses and use of non-DCE preference methods (e.g., Best-Worst Scaling Types 1 - 3, Swing Weighting, Probabilistic Threshold Technique).	24	3	0	7	-12	23	35	116	-300	140
16	Revealed preferences - role in decision making	When and how might revealed preferences be used for decision making in the medical product lifecycle?	24	4	1	5	-14	25	39	116	-350	150

item	Levels	Long	Important	Done already	Too difficult	Not important	B/W	Best	Worst	times	sum	rank
6	Educational Materials - Which Material to Enhance?	What information (e.g., risk information, disease context) benefits most from the use of enhanced educational material (such as videos, voiceovers, gamification, and animations) to inform patients?	23	7	1	3	-17	17	34	116	-425	160
7	Educational Materials - Digital v. Text Formats	How do different types of enhanced educational material (such as videos, voiceovers, gamification, and animations) affect engagement, understanding, choice consistency, and preferences compared to static text and images?	24	6	1	3	-30	13	43	116	-750	170
11	Psychological Constructs - explain preferences across methods	To what extent are relationships between measures of psychological constructs and patient preferences consistent across preference elicitation methodologies (e.g., are relationships between psychological constructs and preferences found with a DCE similar to the relationships found between psychological constructs with Best-Worst Scaling?)?	23	2	3	6	-43	12	55	116	-1075	180

Appendix Table A3. Stakeholders Outside Academia BWS/Importance Results

item	Levels	long	Important	Done already	Too difficult	Not important	B/W	Best	Worst	times	sum	rank
5	Method selection guidance	How to determine which preference assessment method to use in a given context, patient population or for a specific research purpose?	26	9	3	4	40	62	22	148	1000	10
9	Internal Validity / Data Quality	How should one best assess whether patients understand and are paying attention to a given set of cognitive tasks?	22	14	3	3	29	48	19	148	725	20
14	Synthesis of Preferences Across Studies	How to best synthesize multiple patient preference studies for either meta-analysis or predicting preferences for a particular context?	31	4	5	2	26	51	25	148	650	30
3	Attribute presentation & framing	How do changes in the framing (e.g., mortality vs. survival) and attribute presentation (e.g., graphical representation of risk versus text) impact results for a given method?	26	12	2	2	20	37	17	148	500	45
15	Mapping patient-reported outcomes to preference study attributes	How can attributes in a patient preference study be mapped to patient-reported outcomes (or clinical outcome assessments in general)? (e.g., in mapping preferences to a patient-reported outcome in a clinical trial, or incorporating a patient-reported outcome within a preference study)?	33	3	1	5	20	44	24	148	500	45
12	Changes in preferences over time	Which factors influence the stability of preferences over time and why? (e.g., changes in health states, adjustment to condition, nature of illness and treatment, and changes in knowledge)	25	4	10	3	19	44	25	148	475	60
1	Comparing methods	How do the preference mean results and preference heterogeneity results of different patient preference methods compare when applied to address the same research question using the same attributes and samples from the same population?	32	4	5	1	17	57	40	148	425	70

item	Levels	long	Important	Done already	Too difficult	Not important	B/W	Best	Worst	times	sum	rank
19	Study non-DCE Methods	Develop evidence-based good-research practices on the conduct, analyses and use of non-DCE preference methods (e.g., Best-Worst Scaling Types 1 - 3, Swing Weighting, Probabilistic Threshold Technique).	34	5	1	2	13	48	35	148	325	80
18	Expressing uncertainty in patient preference studies	When and how should uncertainty around benefit and risk estimates be incorporated into the design of patient preference studies?	30	6	6	0	11	40	29	148	275	90
4	Transferability across populations or related diseases	How transferable are preferences from one specific disease population to another population (e.g., related diseases, different diseases but similar complaints, same disease but different countries)	27	3	10	2	-2	43	45	148	-50	100
8	Educational Materials - Low Literacy and Numeracy	What types of educational materials are optimal for samples where low literacy and/or low numeracy may be prevalent?	29	7	1	5	-3	28	31	148	-75	110
2	Changing number of attributes	How do changes in the number or types (e.g., categorical vs. numerical value) of attributes impact results for a given method?	27	6	2	7	-5	35	40	148	-125	120
10	Psychological Constructs - explain preferences/ heterogeneity	In which situations do psychosocial constructs (e.g., personal beliefs/personality traits or attitudes) have value in explaining preferences and preference heterogeneity?	20	9	6	7	-10	21	31	148	-250	130
13	Individual preferences	How can individual preferences be used in shared decision-making, (e.g., in the development of decision aids or value clarification)?	26	6	3	7	-12	36	48	148	-300	140
17	Revealed preferences - external validity	How well do stated preferences match revealed preferences in different disease areas or health care decisions, and under what conditions would we expect them to differ?	28	2	5	7	-18	30	48	148	-450	150

item	Levels	long	Important	Done already	Too difficult	Not important	B/W	Best	Worst	times	sum	rank
6	Educational Materials - Which Material to Enhance?	What information (e.g., risk information, disease context) benefits most from the use of enhanced educational material (such as videos, voiceovers, gamification, and animations) to inform patients?	23	9	3	7	-23	21	44	148	-575	160
7	Educational Materials - Digital v. Text Formats	How do different types of enhanced educational material (such as videos, voiceovers, gamification, and animations) affect engagement, understanding, choice consistency, and preferences compared to static text and images?	24	7	5	6	-38	18	56	148	-950	170
11	Psychological Constructs - explain preferences across methods	To what extent are relationships between measures of psychological constructs and patient preferences consistent across preference elicitation methodologies (e.g., are relationships between psychological constructs and preferences found with a DCE similar to the relationships found between psychological constructs with Best-Worst Scaling?)?	20	2	10	10	-40	14	54	148	-1000	180
16	Revealed preferences - role in decision making	When and how might revealed preferences be used for decision making in the medical product lifecycle?	26	5	3	8	-44	26	70	148	-1100	190

Chapter 10

Discussion

Twice during my doctoral research (**Chapter 1** and **Chapter 8**), we assessed the priorities of stakeholders in the field to identify methodological questions regarding the assessment of patient preferences that are of concern to stakeholders who would potentially use this information in decision-making along the Medical Product Lifecycle (MPLC). While the priorities of stakeholders have shifted somewhat since my research (and the IMI-PREFER project) began, they remain focused on practical questions which underly the validity of the preference assessment as well as questions about ways to apply patient preference assessments more broadly (**Chapter 8**). The need to provide evidence to answer these questions has never been more relevant as there are increasing calls for the inclusion of patient preferences in decision making [1-3]. In this general discussion, the thesis aims will be revisited by discussing the main findings for the methodological questions along with evidence that I generated to help answer these questions. This will be followed by a discussion of the questions and future research areas of research that I believe to be of high priority.

COMPARISON OF METHODS

In the initial prioritization survey (**Chapter 1**), the highest prioritized question was comparing the results of simpler/cheaper methods to more complex/expensive methods. This topic was again highly prioritized in **Chapter 8** to increase acceptance of preference methods and their results by decision makers. While there is no 'gold standard' method for patient preference assessment, one method has established itself as a frontrunner for that label. The discrete choice experiment (DCE) has a relatively long history of research supporting its validity and is widely accepted making it often the first choice when researchers want to assess patient preferences [4, 5]. The validity of DCEs is connected to the long and thorough development and analysis process which results in a lot of specific information regarding exactly how important the attributes of a medical treatment are and how important changes in these attributes are relative to each other. This highly detailed information can be used for many different purposes including identifying whether patients would be willing to pay for new treatments or whether new products have acceptable benefit-risk profiles [5]. It is not for nothing that DCEs are one of the most used methods to assess patient preferences [4]. While highly robust, DCEs also require large investments of time and resources to design, implement, and analyze the study outcomes compared to other preference methods. These resources can be prohibitive to the assessment of patient preferences in instances when time is limited, for example when information is needed quickly to inform clinical outcomes in early phase clinical trials, or when only small sample sizes are available.). There is a recognized need to expand the toolbox of instruments to measure patient preferences which can be applied in situations where a DCE is either not needed or not feasible. While many tools currently exist to address this need [6], they may not be as familiar as DCEs and often do not come with a wealth of literature supporting their usage raising questions about their validity and how well they compare to DCEs. This issue was addressed in **Chapter 5** and **Chapter 6** where

Swing Weighting with Direct Rating was compared to DCEs in two case studies assessing patient preferences for glucose monitoring in the self-management of diabetes and treatment preferences of patients with non-small cell lung cancer. A Swing Weighting with Direct Rating task (SW-DR) can be quickly developed and analyzed as it does not require complex designs and directly elicits preference information rather than having that information come from the result of econometric modelling. For this reason it has been identified as a promising preference instrument [7]. In both studies, the outcomes of the Swing Weighting differed from those of the DCE in how important respondents said the attributes were relative to each other. Based on the results of the DCE, there were clearly attributes which were more important or less important for the respondents when making a choice whereas in the SW-DR the attributes were more similar in their importance. This may have been a result of the non-forced tradeoff design of the direct rating task where respondents were not restricted in how many points they could give to each attribute. This is different than in a DCE where you are forced to choose between alternatives sometimes meaning that you need to sacrifice gains in one attribute for gains in another even if both attributes are important. This raised a question about whether the difference between the two methods was related to the non-forced tradeoff in a SW-DR. Therefore, a small, exploratory study was also done as a part of the study in **Chapter 6** to see if a different type of rating scale (i.e., point allocation) would more closely resemble the outcomes of a DCE. In this task, respondents had a limited number of points to allocate to the different attributes meaning they had to make sacrifices like in a DCE. The results more closely resembled DCEs regarding the relative importance of the different attributes, but there was some concern that one attribute (5-year survival) was now even more dominant indicating that this method may be susceptible to over-weighting of dominant attributes. Swing weighting using point allocation for the direct elicitation of attribute weights is a promising method which is quick and easy to implement [8], but more research is needed to understand the factors which may influence over- or under-weighting of attributes and the use cases where each method may be more suitable.

Another primary use of DCEs is understanding the tradeoffs that patients would be willing to make in their care. Multidimensional-thresholding is a method that expands on probabilistic thresholding which can generate trade-off information and part worth utilities across multiple attributes and levels by first identifying the rank of different attributes and then identifying the threshold where someone would trade a benefit for a risk [9-11]. In this way it produces similar information to that of a DCE but without the sample size or experimental design requirements as the information is directly elicited. This makes its application and usage much quicker, but the method is relatively new and unknown and therefore not often utilized. Like Swing Weighting, it can be used to produce similar outcomes to DCEs but lacks the long history of literature supporting their theoretical basis and use in healthcare decision making. This lack of supporting literature hinders uptake as researchers may be unaware of these tools or may be hesitant to use tools which may not be accepted by stakeholders who use this information.

This results in a vicious cycle inhibiting innovation as the lack of evidence supporting their acceptance results in hesitation to use these techniques until more evidence supporting their acceptance is available, but this evidence can only be generated when more researchers use these techniques. Collaborative research initiatives (like the IMI-PREFER project [12]) may need to lead the way in generating this evidence. These collaborations not only allow multiple stakeholders to share the burden of the costs, but also ensure that the evidence generated meets the needs and concerns of these different stakeholders.

ATTRIBUTE FRAMING

Another question that was identified as being a priority in **Chapter 1** was the impact of attribute framing on preference outcomes. Attribute development and determining the way that the attributes are framed are two important steps in developing patient preference assessment tools as the attributes can determine whether or not the preference information is applicable and the way the attributes are framed can influence the outcomes of the study [13, 14]. While research has been done looking at how to present risk information [15-18], there are still gaps in knowledge regarding the impact of attribute framing on other attributes like whether a treatment works [19]. In **Chapter 4,** I presented the outcomes of a case study in which participants were asked to complete a discrete choice experiment which varied in the way the efficacy of an antibiotic treatment was framed (Effectiveness Rate, Failure Rate, or both). We found that while attribute framing was associated with different preference outcomes, these differences were small and did not change our understanding of what is important to patients but may be significant in other ratio-based assessments like willingness to pay or maximum acceptable risk.

This study adds to our understanding of how attribute framing has an impact on preference outcomes which has implications for the design of preference experiments. The difficult thing with framing is that there is not necessarily a "right way" to frame attributes [19] and our study showed that presenting multiple frames may have issues on its own. Howard and Salkeld suggested adding a step during attribute development to assess whether there was a risk of framing bias for the attributes and levels [20]. While this may not help to answer the question of how to frame an attribute, it at least provides a point in which researchers can stop and consider the decision-making contexts in which the patient preference information will be used and how it is framed there. Researchers should also look at how attributes are framed in clinical practice and in previous preference studies to try and align their definitions with those. A library of attributes and levels was identified as desirable by stakeholders in the field of patient preference assessment in **Chapter 8**. Such a library will not only help to align attribute framing but will reduce the proliferation of attributes with slightly different definitions used to measure the same underlying construct which can impact transferability. Ultimately, aligning the framing with how it is already used may not reduce the risk of a framing effect but will

at least reduce the risk of introducing a bias to the interpretation of preference outcomes, increasing their relevance and support their use by stakeholders.

TRANSFERABILITY

While ideally the patient perspective would be included at every step during the MPLC, the reality is that patient preference information is often categorized as only 'nice to have' information with time and resources for new research only coming when the information has been specifically requested (i.e. by regulators [21-23]). One major reason for these resource allocation restrictions is that patient preference studies are often bespoke and developed in relation to a specific treatment profile. While this is an optimal approach, it may not always be necessary if the results of previously conducted preference studies can be transferred to new decision contexts. For this reason, the topic of transferability of patient preferences was one of the most prioritized research topics in both **Chapter 1** and **Chapter 8**.

One of the major outcomes of the IMI-PREFER project is a checklist of relevant information which can be used to support the transfer of patient preference information to a new decision context, either being another country or another application of the same drug in a different patient population or for another indication. Building upon previous work in health economics [24], the transferability checklist consists of 3 groups of 13 characteristics. These groups are related to whether the preference task is applicable to this new context (Methodological characteristics), whether the sample populations are comparable and if aspects of the demographic profile or attitudes and beliefs of these samples would result in different preferences (Population characteristics), and whether the healthcare context in which the preference is elicited would influence the preference outcomes (see Figure 1)[25]. Evidence related to these characteristics can be used to support the transferability of preference outcomes.

Three times during my research we compared the preferences of different groups of respondents. In **Chapter 2** we compared the preferences of diabetes patients in the Netherlands and Poland for glucose monitoring technology. We found that preferences differed not only between the country samples, but also between different patient sub-groups in each country (such as preferences of those aged 18-50, whether the patients currently use fingerpicking to measure blood glucose, and the type of diabetes they have). When assessing the transferability of these results using the checklist, we found that the primary limitation to transferability would be the different healthcare contexts and specifically how these glucose monitors were reimbursed in different locations. If these healthcare contexts were similar in the Netherlands and Poland, then the results would probably be transferable. In **Chapter 3** we compared common opinions about the most important areas of unmet medical needs in patients with two types of neuromuscular diseases (myotonic dystrophy type 1 and mitochondrial disorders) and their

Figure 1. PREFER Transferability Checklist

Methodological characteristics

- attributes
- levels and their range

Population characteristics

- sociodemographic and educational characteristics of the population
- epidemiologic characteristics
- attitudinal characteristics (only for transferability across countries)
- cultural and religious beliefs (only for transferability across countries)
- cognitive characteristics (only for transferability across diseases / patient groups / indications)

Healthcare context characteristics

- commercial and financial (reimbursement) availability of medical products
- geographical accessibility of medical products
- level of experience with use of medical products or adverse health effects
- level of trust of patients in treatment with a medical product
- the standard of care for treatment of a certain disease
- healthcare and social security system (only for transferability across countries)

caregivers. We found differences in opinion not only between the patients and caregivers but also between patients with different types of neuromuscular diseases. These differences were most pronounced in the smaller, more unique sub-groups. Transferability of these preferences would likely be limited to other neuromuscular disease with different epidemiological characteristics, clinical symptoms, and disease severities. In an additional case study (not presented in this thesis), we looked at the differences in preferences for the treatment of lung cancer in Belgium and Italy using a DCE [26]. Here a major issue was how dominant the rating of 5-year survival was in both countries and therefore the high levels of the side-effect risk patients would accept to increase this survival. We would thus expect the results to be generally transferable when transferred to other cancer types with similar mortality rates. Interestingly, the countries did differ in their preferences for how the treatment was administered (oral versus infusion in a hospital setting). The Italian population significantly preferred oral administration over infusion while the Belgian sample did not significantly prefer

one administration over another, possibly illustrative of differences in geographic accessibility of medical care.

In these three cases, the methodological contexts, the sociodemographic/educational and epidemiological characteristics, and the healthcare contexts related to costs were the primary factors when deciding if outcomes were transferable. Other case studies in PREFER also assessed which factors would limit or support the transferability of their results using the checklist. The above mentioned characteristics along with the healthcare contexts of healthcare access, disease history, and treatment familiarity were the main aspects discussed when considering the transferability of preferences [25]. Other aspects like attitudinal characteristics, or cultural and religious beliefs were sometimes mentioned but only when they were either related to a specific research question being investigated or when the researchers responsible for the case study hypothesized that these may also be relevant but were not investigated.

The checklist provides a good overview of concepts to consider when evaluating the transferability of preferences from one context to another. In the end, the final assessment of whether preference information can be transferred to new decision contexts remains a qualitative and subjective assessment for which more information can always be used. Much of the information that PREFER researchers used to assess transferability of their cases studies is commonly included in publications (the attributes and levels; sociodemographic/educational and epidemiological characteristics), but much of the additional information identified in the checklist is not (healthcare contexts related to costs, access, disease history, treatment familiarity, and level of trust in treatment; attitudinal characteristics; cultural and religious beliefs; cognitive characteristics). This means that even if preferences would be suitable to be transferred to a new decision context there may not be enough information available to support the transferability.

One possible reason for the lack of information to support the transferability of outcomes may be that researchers are restricted in how much information they can publish. Just because information is not available in published or publicly available information does not mean that this information does not exist or was not collected. Restrictions on the number of words and tables/graphics often forces researchers to selectively choose what information is reported. Providing additional information relevant to the characteristics in the transferability checklist as supplementary material is one possible way to ensure that the current standards in academic publishing are maintained while allowing for greater publication of available data. If stakeholders who use preference information would like to see an increase in the reuse of preferences, then transferability should be a topic of further discussion in professional organizations where patient preference research is done. These organizations can provide guidance on additional information that should be reported in these areas and possibly provide support for a repository of this information. Without this, the only option left to researchers

is reaching out to the corresponding author of an article who may or may not be reachable, and who may or may not be able to access the necessary information.

Finally, it may not always be the case that more information is needed to transfer preference outcomes. Greater levels of information are needed to support the transferring of preference information to more impactful decision points in the medical product lifecycle (such as identifying target product profiles or approving market access [27]), but other uses may be more amenable to lower levels of supporting information. Instances such as the identification of concepts of interest to support the selection of clinical outcome assessment or patient reported outcome measures may not require as much supporting evidence when reusing preference outcomes from other disease areas or treatments [28, 29]. Future research could look at what levels of information stakeholders would require to support the use of transferred preference outcomes to different decision-making contexts as some use cases may be less restrictive than others thus lowering the barriers to a wider application of preference information.

EDUCATIONAL TOOLS: WHAT SHOULD WE BE LOOKING AT NOW?

One topic which I spent much of my PhD working on was the use of video-based information to educate patients prior to completing a choice task. The aim of these educational tools is to ensure that the sample completing a preference task is informed when giving their preferences to support the validity of the findings. Uninformed preferences could undermine the validity of preference outcomes as respondents may have reported different preferences if they better understood the task and attributes. Informing respondents using educational information can reduce the chance of receiving uninformed responses. Questions remain about the best format to provide this information in. Video-based educational tools are one potential option. However, it is not clear based on the research presented in **Chapter 7** that simply providing information in a video succeeds at informing patients and previously published studies are mixed on the impact or practical relevance of using video-based information [30-33]. In **Chapter 8**, we asked if this was still a prioritized research question. It was found to be a much lower priority than other topics with a large portion of respondents saying that it had been studied sufficiently. In response to this, I would have to say that I agree. While there may be a role for video-based information (such as for those with low levels of literacy or people with special needs like those who are blind or may suffer from cognitive impairment related to their illness - as was the case for some patients in the study discussed in **Chapter 3**, other questions related to education materials may be more important to answer to help with future preference studies.

Specifically, if the aim of this information is to ensure respondents are informed to support the validity of the preference outcomes [34, 35] then we first need to answer fundamental questions related to what a sufficiently informed respondent is, how researchers can check if respondents

are informed, what to do with uninformed preferences, and how to get respondents to engage with educational material. The first issue is difficult, and the answer (when given by researchers) is often formulated in a vague manner. We defined an informed respondent using a standard definition as one who understood the attributes, levels, and choice task as we intended them [36], but we did not specify exactly what we meant by saying "how we intended them to". Others have used similarly vague definitions or use no definitions at all and say simply that they tested understanding using rationality tests [35]. Clarifying what information is needed to be sufficiently informed is needed prior to check whether respondents are informed. Further complicating this matter is the issue that the information presented in a preference study is often complex and may require different types of understanding.

The example of risk information is an excellent example as it is both commonly used and notoriously difficult for patients to understand [15, 18, 19, 37]. In **Chapter 2** we assessed preferences for risk of skin irritation at four different levels: 0% risk, 5% risk, 20% risk, and 35% risk. We presented this information in three different ways according to best practices (absolute risk, proportional risk, and using icon arrays) [15]. To fully understand this risk information a respondent would need to know that an increase from 5% to 20% is a) a 15 percentage-point increase, b) a relative increase of 300%, and c) the patient has a 1 in 5 chance of having skin irritation, which is d) four times higher risk than at 5% when it was 1 out of 20. This attribute level change also needs to be understood in relation to the increase from 20% to 35% which has similar implications (15 percentage-point increase) and also very different ones (7 in 20 chance of having skin irritation; relative increase of 75%). Further complicating this issue is that patients may use decision heuristics to simplify the risk information during a choice task [38]. One relevant heuristic, ordinal recoding, involves converting numerical levels into qualitative levels for simpler comprehension [39]. In the example above that may mean patients would convert 0%, 5%, 20%, and 35% risk to lowest, low, moderate, and high risk. When this happens it threatens the validity of more complicated applications of preference data (such as the calculation of marginal rates of substitution where point estimates are needed) [38]. Thus, to define what a sufficiently informed respondent is, researchers first need to identify the use case for the preference outcomes and define the level of understanding necessary to be sufficiently informed when completing the preference tasks. Once this definition is available researchers can progress to the second aspect of checking whether respondents are at that level.

One common method used to support preference outcomes as coming from informed respondents is the use of comprehension checks [35]. Comprehension checks are objective questions meant to directly test if respondents understand specific information. These checks play a dual role to both support the validity of the preference outcomes as well as providing a teaching moment to correct respondents who may not understand specific information or survey elements [40]. However, these checks often consist of one-time, individual questions with correct answers presented after the question. These do not assess whether the correct

answer was able to educate the respondent or whether the successful answering of these comprehension tasks means that the respondent was sufficiently informed. In instances where a deeper understanding of the information is needed (as defined by the researchers), follow up comprehension checks could be used to verify if the respondent was sufficiently informed. However, researchers should be judicious in their application of extra checks as respondents may feel that they are being tested if too many are used, resulting in higher levels of dropouts. Other methods such as thorough qualitative pre-testing of surveys and the inclusion of educational materials are commonly mentioned to support respondent understanding [35]. While both can support a better understanding of the preference task neither ensures that respondents are informed as the settings of the pre-test and the actual survey may not be comparable [35, 41] and the provision of information does not always equate to engagement with information (see **Chapter 7**). Thus, while both should still be done, they cannot be used as evidence to support the preferences as being informed.

The question then arises about what to do with responses that were found to come from possibly uninformed respondents. As these preferences can undermine the validity of the preference outcomes it would be reasonable to exclude these responses from the final analysis to ensure that the outcomes reflect what patients would want in their care. However, uninformed preferences may still reflect the values of a portion of the general patient population for which the healthcare decisions are relevant as not all patients in the general population are well informed either. Removing this group of respondents from the final analysis threatens the generalizability of the preference outcomes to the wider patient population. So, for preference outcomes to be both valid and generalizable the outcomes from both informed and uninformed respondents need to be considered. Sensitivity analyses which assess the preferences of both groups individually and combined can identify whether the validity of the preference outcomes is at risk due to uniformed responses while the combined analysis should reflect the preferences of the general patient population. Presenting these analyses together will help to support the validity and generalizability of the preference outcomes while addressing the weakness of presenting one analysis alone.

Finally, while there is a place for the inclusion of uninformed preferences in decision making, the primary goal of preference assessment is to understand the preferences of informed patients. Educating respondents prior to the completion of a preference task remains the best way to reach this goal. There is a wealth of information available on the most effective ways to educate patients when making a decision [42], but none of that matters if the respondent does not engage with the materials (**Chapter 7**). Identifying ways to increase engagement with educational materials will not only increase the number of informed respondents in the sample but may also increase engagement with the survey in general increasing the data quality and validity of the outcomes. Cheap talk (asking respondents to pay more attention by explaining the issues that arise when they do not pay attention) has been presented as a

way to increase engagement with the choice task [43], but has not been investigated as a way to increase engagement with educational material to my knowledge. Finding ways to increase engagement with educational materials should be a priority topic of research in the future.

POTENTIAL ISSUES WITH CALCULATING WILLINGNESS-TO-PAY

As a primary outcome of the diabetes case study (**Chapter 2**), the willingness-to-pay (WTP) estimates for different glucose monitoring devices were calculated based on the outcomes of a discrete choice experiment. WTP estimates are a common outcome of preference studies used to assess what monetary value respondents put on different levels of attributes as a way to quantify value using a universal and easy to understand unit of measurement [44]. WTP estimates can be derived by including a cost attribute in discrete choice experiments. Part-worth utilities per monetary unit can then be derived using the coefficients of the final econometric model (e.g. the value of a euro to a respondent) [45]. The same models can also be used to calculate the utility difference between attribute levels or the utility of a medical product profile. WTP estimates can then be generated by dividing the utility of a healthcare profile or utility difference between attribute levels by the negative utility per monetary unit. The resulting number represents the highest cost level at which a respondent would theoretically be willing to pay for a medical product or change in a product.

Underlying the economic theory of these models is the assumption that the respondent attended to and considered the relevant attributes when making their choice. When this is not the case and respondents make their choice without regard to the price or other relevant attributes, standard modelling techniques are not suitable as they will produce coefficients as if they had been fully considered [46]. This can result in inaccurate or invalid ratio-based outcomes like WTP, maximum acceptable risk, or minimum acceptable benefit [47].

Recognition of this problem is not new. Hensher, Rose & Greene reported on a study assessing car commuter trips in Sydney where they tried to account for this using self-reported non-attendance to attributes [48]. They found that WTP estimates were significantly lower when 'tailoring' the WTP estimation process by excluding attributes based on respondent reports on ignoring those attributes in making choices lower (18-62% depending on the attribute). Similarly, Carlsson, Kataria & Lampi conducted a study looking at how the WTP of people living in Sweden for three different environmental objectives changes when the estimates are restricted to 0 for non-attenders [49]. Their WTP estimates were significantly lower when they accounted for attribute non-attendance using self-reports of attendance in their analysis. Scarpa et al. looked at this issue in a study of the general public's attitudes and preferences regarding rural environmental landscape improvements in the Republic of Ireland [46]. They used a latent class analysis (LCA) to create models excluding respondents who likely did not attend to different attributes. They found that Cost was one of the most non-attended to

attributes and accounting for potential non-attendance to cost resulted in WTP estimates which were much lower and more "realistic." In the field of healthcare, Lagarde identified this problem in her study of a DCE assessing preferences for the introduction of new guidelines to manage malaria in pregnancy in Ghana [50]. In this DCE, Lagarde included a cost attribute related to monthly bonuses to be paid to healthcare personnel in these clinics. Lagarde used a latent class analysis (LCA) to create seven models to account for attribute non-attendance in the respondents: one model for respondents who used all the attributes in decision making and six to identify respondents who ignored one of the attributes in the DCE. She found that when non-attendance was accounted for, respondents thought monthly bonuses for healthcare personnel working in malarial pregnancy clinics could be up to 11.65 times higher than when cost non-attendance was not accounted for.

To illustrate this potential issue when generating WTP estimates, I used the data from **Chapter 2** to look at how monthly out-of-pocket WTP estimates change when calculated using different methods. The specific monthly out-of-pocket WTP assessed were the estimates to change a glucose monitoring device from requiring 2 fingerpricks per day to a device with 0 fingerpricks per day. The estimates were calculated in four different ways. First, the main model coefficients from the mixed effects model presented in **Chapter 2** were used to calculate WTP estimates as this is the standard practice in the field (Method 1). Then the individual coefficients generated in the mixed effects model were used to calculate the WTP estimates per respondent. The individual WTP estimates were used to identify respondents who would not be willing to pay (WTP estimates less than €0) for this change and respondents who had WTP estimates that were above the price levels used in the DCE (€250). These individual WTPs were then either removed from the analysis (censored) if their WTP estimates were above €250 or less than €0 (Method 2) or they were adjusted up to €0 or down to €250 if their WTP estimates were less than €0 or above €250 (Method 3) [51]. Finally, the latent class analysis approach as described by Lagarde [50] and Scarpa et al. [46] was used to identify respondents who attended to the attributes of concern using a 5-class model (Method 4). In this method, the five model classes were 1: attendance to all relevant attributes, 2: non-attendance to all attributes, 3: non-attendance to fingerpricks, 4: non-attendance to costs, 5: non-attendance to fingerpricks and costs. Models 2 through 5 have forced coefficient parameters = 0 for the attributes named to replicate non-attendance to those attributes. The coefficients from model 1 are used to generate the WTP estimates as they should represent the respondents who attended to both relevant attributes and thus fulfill the assumptions needed for these calculations.

As can be seen in Table 1, large differences in WTP estimates were found when using the different methods. The WTP estimates from the latent class model of respondents who attended to both the number of fingerpricks and costs had the lowest WTP (€29.90) while the mean of the censored individual WTP estimates had the highest (€58.06). There was a 42% reduction in WTP estimates between Method 4 (LCA accounting for non-attendance)

and Method 1 (standard method) which is in line with previous findings. The individual WTP estimates ranged from -1429 to 262,379 with 13.7% of respondents having a WTP less than 0 indicating that they would not pay at all for the change in number of finger pricks. In addition, the class membership estimates from the latent class analysis indicate that 49.2% of respondents likely fell in a class that did not attend to costs and thus violated the assumptions needed to be included in the analysis (either because they did not attend to the cost attribute (Model 4: 31.6%), they did not attend to both costs or fingerpricking (Model 5: 16.8%), or they did not attend to any attributes at all (Model 2: 0.8%)).

Table 1. Results of sensitivity analysis of monthly out-of-pocket WTP estimates for reducing fingerpricks from 2 to 0, estimated using four different methods to account for attribute non-attendance

	Method 1: Total Sample MIXL	**Method 2: Censored individual WTP estimates**	**Method 3: Adjusted individual WTP estimates**	**Method 4: LCA: Model 1: Attendance to relevant attributes**
WTP	42.46	Mean = 58.06 (SD=70.91)	Mean 48.34 (SD=44.31)	29.90

Note: Utility was calculated as the difference in model coefficients between the two levels of fingerpricking or the difference for a one unit increase in costs. WTP: Willingness to pay; SD: Standard Deviation; MIXL: Mixed effects logit model; LCA: Latent Class Analysis

The question then is which method to use when generating WTP estimates. Both Hensher, Rose & Greene and Carlsson, Kataria & Lampi used self-reports of non-attendance to identify respondents to be excluded. But self-reported non-attendance has previously been found to not be a reliable way to identify non-attenders [52]. Carlsson, Kataria & Lampi even reported that there were instances where inferred non-attendance (non-attendance identified using statistical methods) was not in concordance with the respondents' stated non-attendance. Using the individual WTP estimates is highly susceptible to outliers and there is a question of how to calculate accurate estimates. Lagarde and Scarpa et al. propose using statistical methods to isolate those who did not attend to the cost attribute according to the model coefficients. However, coefficients of 0 (or not significantly different from 0) do not necessarily mean that the respondent did not attend to the cost attribute. These coefficients could also occur when a respondent carefully reviews the attribute levels in a DCE profile and judges that the attribute is simply not important to their decision making in that context.

The only conclusion that I can draw from this exercise is that there is no "correct" way to account for non-attendance when calculating WTP estimates or any ratio-based estimates (such as maximum acceptable risk [47]) for that matter. Sensitivity analyses like this are important to support these outcomes and can provide additional relevant information to help decision makers when using this information. Especially the secondary information showing the proportions of respondents who may not have attended to an attribute or who would not be willing to pay for improvements like this is relevant to help understand the

heterogeneity of preferences. This issue is not something that I have seen often addressed by researchers who calculate these types of outcomes. As Lagarde says "inferring anything about the willingness to pay of all respondents is misleading, and researchers should try to reflect better the heterogeneity of valuations" [50]. Researchers should be aware of potential violations of the assumptions underlying the models they use to calculate these estimates and conduct sensitivity analyses to identify uncertainties with these outcomes. This can help ensure that when decision makers use this information the decisions they make are well supported.

CONCLUSION: THOUGHTS ON WHAT SHOULD BE PRIORITIZED FOR FUTURE RESEARCH IN PATIENT PREFERENCE ASSESSMENT

While much research has been done in the world of patient preference research it is still a relatively young field. As a conclusion to this thesis, I would like to highlight areas where I think that future research into patient preference should be prioritized. The first area that I think should be prioritized for future research is looking at the stability of preferences. This was a research topic that was listed as being high priority in **Chapter 8**, but also one that was also listed as being too complicated or impossible to study. We often only measure preferences at one fixed time and then apply this information as if preferences are static. Evidence is mixed on how stable preferences are with some research finding that preferences are stable and other research not [53-57]. As care evolves and patients live longer with more chronic diseases, understanding preferences before and after treatment or over time as disease experience changes will become increasingly relevant not only to medical product development but to clinical practice in general [58]. This is especially relevant in cases where we ask respondents to give their opinion on hypothetical situations where they may not have personal experience with the side effects or low efficacy rates of treatment as references to guide their preferences [59, 60]. Patient centric decision making in the medical product lifecycle requires that we make the patient central to the decision. If the patient is dynamic and changing, then our understanding of their preferences should be dynamic and change with them as well.

This thesis presented evidence to help support the increased acceptance of preference methods and their results by decision makers. The research done consisted primarily of empirical research to address methodological questions to support the validity of preference assessment. The assumption here is that if valid preference information is available then it should be used in decision making. If I were to say what area I think should be prioritized for research moving forward to increase the use of preference information in decision making I would take a more practical approach. Specifically, future research should focus on collaborations between regulatory or government agencies and preference researchers to generate evidence related to the application of preference information by these decision-making bodies.

Underlying all the work done in this thesis is the primary aim that preference research will be used to inform decision-making. If preference information does not meet the needs of decision makers, then it will not be used for these purposes. In two of the sections included in this discussion (Comparison of methods, Transferability) not meeting the needs of decision makers was named as a potential barrier to the use of patient preference information. But it is not clear exactly what the needs or standards would be for preference information to be useful. Further complicating this is the recognition by regulatory agencies that the applicability of patient preference information is scenario-dependent and that their experience in using this information is currently limited [61]. Research collaborations can help to generate evidence on how to meet these needs and provide evidence for the successful use of preference information (or at least specify why this did not happen). Pushing for more collaborations can help speed up the generation of evidence which will likely increase the assessment of preferences and their use in decision making.

REFERENCES

1. Jackson, Y., et al., The evolving role of patient preference studies in health-care decision-making, from clinical drug development to clinical care management. Expert Review of Pharmacoeconomics & Outcomes Research, 2019. **19**(4): p. 383-396.
2. Shimeng, L., et al., The Value, Methods and Challenges of Patient Preference Evidence Applied to Health Technology Assessment. Chinese Health Policy Research, 2022. **15**(5): p. 52-58.
3. Bouvy, J.C., et al., Use of Patient Preference Studies in HTA Decision Making: A NICE Perspective. The Patient - Patient-Centered Outcomes Research, 2020. **13**(2): p. 145-149.
4. Soekhai, V., et al., Discrete Choice Experiments in Health Economics: Past, Present and Future. Pharmacoeconomics, 2019. **37**(2): p. 201-226.
5. Ryan, M., K. Gerard, and M. Amaya-Amaya, Using discrete choice experiments to value health and health care. Vol. 11. 2007: Springer Science & Business Media.
6. Soekhai, V., et al., Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discovery Today, 2019. **24**(7): p. 1324-1331.
7. Whichello, C., et al., Appraising patient preference methods for decision-making in the medical product lifecycle: an empirical comparison. BMC Medical Informatics and Decision Making, 2020. **20**(1): p. 114.
8. Schneider, P.P., et al., The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states. Wellcome Open Res, 2022. **7**: p. 14.
9. Hauber, B. and J. Coulter, Using the Threshold Technique to Elicit Patient Preferences: An Introduction to the Method and an Overview of Existing Empirical Applications. Appl Health Econ Health Policy, 2020. **18**(1): p. 31-46.
10. Postmus, D., et al., Individual Trade-Offs Between Possible Benefits and Risks of Cancer Treatments: Results from a Stated Preference Study with Patients with Multiple Myeloma. Oncologist, 2018. **23**(1): p. 44-51.
11. Postmus, D., et al., Incorporating patient preferences into drug development and regulatory decision making: results from a quantitative pilot study with cancer patients, carers, and regulators. Clinical Pharmacology & Therapeutics, 2016. **99**(5): p. 548-554.
12. de Bekker-Grob, E.W., et al., Giving patients' preferences a voice in medical treatment life cycle: the PREFER public–private project. 2017, Springer. p. 263-266.
13. Lancsar, E. and J. Louviere, Conducting Discrete Choice Experiments to Inform Healthcare Decision Making. PharmacoEconomics, 2008. **26**(8): p. 661-677.
14. Bridges, J.F., et al., Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health, 2011. **14**(4): p. 403-13.
15. Harrison, M., et al., Risk as an attribute in discrete choice experiments: a systematic review of the literature. Patient, 2014. **7**(2): p. 151-70.
16. Veldwijk, J., et al., Survival or Mortality: Framing of the Risk Attribute in a Discrete Choice Experiment. Value Health, 2014. **17**(7): p. A330.
17. Ahmed, H., et al., Communicating risk. BMJ : British Medical Journal, 2012. **344**: p. e3996.
18. Gigerenzer, G., et al., Helping Doctors and Patients Make Sense of Health Statistics. Psychol Sci Public Interest, 2007. **8**(2): p. 53-96.

19. Freeman, A.L.J., How to communicate evidence to patients. Drug and Therapeutics Bulletin, 2019. **57**(8): p. 119-124.

20. Howard, K. and G. Salkeld, Does Attribute Framing in Discrete Choice Experiments Influence Willingness to Pay? Results from a Discrete Choice Experiment in Screening for Colorectal Cancer. Value in Health, 2009. **12**(2): p. 354-363.

21. Ho, M.P., et al., Incorporating patient-preference evidence into regulatory decision making. Surgical Endoscopy, 2015. **29**(10): p. 2984-2993.

22. Janssen, E.M., et al., Identifying and prioritizing concerns associated with prosthetic devices for use in a benefit-risk assessment: a mixed-methods approach. Expert Review of Medical Devices, 2018. **15**(5): p. 385-398.

23. Webber, C.M., et al., Measuring Patient Preferences at the FDA Center for Devices and Radiological Health: Reflections and Projections. Value in Health, 2021. **24**(7): p. 1024-1029.

24. Munthe-Kaas, H., H. Nøkleby, and L. Nguyen, Systematic mapping of checklists for assessing transferability. Systematic Reviews, 2019. **8**(1): p. 22.

25. The PREFER Consortium. PREFER Recommendations: Why, when and how to assess and use patient prefernces in medical product decision-making. 2022 [cited 1st edition]; Available from: https://doi.org/10.5281/zenodo.6470922.

26. Oliveri, S., et al., Balancing benefits and risks in lung cancer therapies: patient preferences for lung cancer treatment alternatives. Frontiers in Psychology, 2023. **14**.

27. Whichello, C., et al., An overview of critical decision-points in the medical product lifecycle: Where to include patient preference information in the decision-making process? Health Policy, 2020. **124**(12): p. 1325-1332.

28. Administration, U.S.F.a.D., Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for Purpose Clinical Outcome Assessments - Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. 2022, FDA, Silver Spring, MD.

29. U.S. Food and Drug Administration, Patient-Focused Drug Development: Methods to Identify What Is Important to Patients Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. 2022, FDA, Silver Springs, MD.

30. Vass, C.M., et al., A picture is worth a thousand words: the role of survey training materials in stated-preference studies. The Patient-Patient-Centered Outcomes Research, 2020. **13**(2): p. 163-173.

31. Lim, S.L., et al., Are Videos or Text Better for Describing Attributes in Stated-Preference Surveys? The Patient - Patient-Centered Outcomes Research, 2020. **13**(4): p. 401-408.

32. Bywall, K.S., et al., Does being exposed to an educational tool influence patient preferences? The influence of an educational tool on patient preferences assessed by a discrete choice experiment. Patient Education and Counseling, 2021. **104**(10): p. 2577-2585.

33. Overbeeke, E., et al., Patient preferences for gene therapy in haemophilia: Results from the PAVING threshold technique survey. Haemophilia, 2021.

34. Janssen, E.M., et al., Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Review of Pharmacoeconomics & Outcomes Research, 2017. **17**(6): p. 531-542.

35. Pearce, A., et al., Respondent Understanding in Discrete Choice Experiments: A Scoping Review. The Patient - Patient-Centered Outcomes Research, 2021. **14**(1): p. 17-53.

36. Helter, T.M. and C.E.H. Boehler, Developing attributes for discrete choice experiments in health: a systematic literature review and case study of alcohol misuse interventions. Journal of substance use, 2016. **21**(6): p. 662-668.

37. Vass, C.M. and K. Payne, Using Discrete Choice Experiments to Inform the Benefit-Risk Assessment of Medicines: Are We Ready Yet? Pharmacoeconomics, 2017. **35**(9): p. 859-866.

38. Veldwijk, J., et al., Taking the Shortcut: Simplifying Heuristics in Discrete Choice Experiments. Patient, 2023.

39. Veldwijk, J., et al., Words or graphics to present a Discrete Choice Experiment: Does it matter? Patient Education and Counseling, 2015. **98**(11): p. 1376-1384.

40. Mansfield, C., et al., Pmu129 Performance Of A Comprehension Question In Discrete-Choice Experiment Surveys (Dce). Value in Health, 2019. **22**: p. S730-S731.

41. Harris, L.R. and G.T. Brown, Mixing interview and questionnaire methods: Practical problems in aligning data. Practical Assessment, Research, and Evaluation, 2010. **15**(1): p. 1.

42. Stacey, D., et al., Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews, 2017(4).

43. Özdemir, S., F.R. Johnson, and A.B. Hauber, Hypothetical bias, cheap talk, and stated willingness to pay for health care. Journal of Health Economics, 2009. **28**(4): p. 894-901.

44. Olsen, J.A. and R.D. Smith, Theory versus practice: a review of 'willingness-to-pay'in health and health care. Health economics, 2001. **10**(1): p. 39-52.

45. Breidert, C., Measuring Willingness-to-Pay, in Estimation of willingness-to-pay: Theory, measurement, application. 2007, Springer Science & Business Media. p. 37-56.

46. Scarpa, R., et al., Modelling attribute non-attendance in choice experiments for rural landscape valuation. European Review of Agricultural Economics, 2009. **36**(2): p. 151-174.

47. Fairchild, A.O., S.D. Reed, and J.M. Gonzalez, Method for Calculating the Simultaneous Maximum Acceptable Risk Threshold (SMART) from Discrete-Choice Experiment Benefit-Risk Studies. Med Decis Making, 2023. **43**(2): p. 227-238.

48. Hensher, D.A., J. Rose, and W.H. Greene, The implications on willingness to pay of respondents ignoring specific attributes. Transportation, 2005. **32**: p. 203-222.

49. Carlsson, F., M. Kataria, and E. Lampi, Dealing with Ignored Attributes in Choice Experiments on Valuation of Sweden's Environmental Quality Objectives. Environmental and Resource Economics, 2010. **47**(1): p. 65-89.

50. Lagarde, M., Investigating attribute non-attendance and its consequences in choice experiments with latent class models. Health economics, 2013. **22**(5): p. 554-567.

51. Chambers, R., et al. Winsorization for identifying and treating outliers in business surveys. in Proceedings of the second international conference on establishment surveys. 2000. American Statistical Association Alexandria, Virginia.

52. Hess, S. and D.A. Hensher, Using conditioning on observed choices to retrieve individual-specific attribute processing strategies. Transportation Research Part B: Methodological, 2010. **44**(6): p. 781-790.

53. Auriemma, C.L., et al., Stability of End-of-Life Preferences: A Systematic Review of the Evidence. JAMA Internal Medicine, 2014. **174**(7): p. 1085-1092.

54. Vina, E.R., et al., Patient preferences for total knee replacement surgery: Relationship to clinical outcomes and stability of patient preferences over 2 years. Seminars in Arthritis and Rheumatism, 2016. **46**(1): p. 27-33.

55. Everhart, M.A. and R.A. Pearlman, Stability of Patient Preferences Regarding Life-Sustaining Treatments. Chest, 1990. **97**(1): p. 159-164.

56. Janssen, D.J.A., et al., Predicting changes in preferences for life-sustaining treatment among patients with advanced chronic organ failure. Chest, 2012. **141**(5): p. 1251-1259.

57. Jansen, S.J.T., et al., Patients' preferences for adjuvant chemotherapy in early-stage breast cancer: is treatment worthwhile? British Journal of Cancer, 2001. **84**(12): p. 1577-1585.

58. Sava, M.G., et al., An analysis of the sensitivity and stability of patients' preferences can lead to more appropriate medical decisions. Annals of Operations Research, 2020. **293**(2): p. 863-901.

59. Hensher, D.A., Hypothetical bias, choice experiments and willingness to pay. Transportation Research Part B: Methodological, 2010. **44**(6): p. 735-752.

60. Harrison, G.W. and E.E. Rutström, Chapter 81 Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods, in Handbook of Experimental Economics Results, C.R. Plott and V.L. Smith, Editors. 2008, Elsevier. p. 752-767.

61. Committee for Medicinal Products for Human Use (CHMP), Qualification Opinion of IMI PREFER. 2022.

General Summary

Patient preferences assessments are defined as qualitative or quantitative assessments of the relative desirability or acceptability to patients of specified alternatives or choices among outcomes or other attributes that differ among alternative health interventions[1]. In simpler terms, these assessments measure what is important or of value to patients in healthcare. The value and usefulness of these assessments is increasingly being recognized as medical product development shifts towards a more patient centric approach. Examples of when patient preference information (PPI) can be used include by industry members to identify promising areas for research or to show that their product has acceptable tradeoffs between risks and benefits. PPI can also be used by regulatory agencies to see if medical products are desired by patients and if patient needs are being met. It can even be used by patient advocates to highlight areas where research should be targeted. These are just a few examples, but in all cases stakeholders who want to use this information want to ensure that the PPI they are using is of sound methodological quality. This thesis aimed to assist in this evaluation by providing empirical evidence related to methodological questions regarding the assessment of patient preference information.

In **Chapter 1** we conducted a survey with members of the PREFER consortium to identify and prioritize methodological questions regarding how to conduct patient preference studies to support their use in decision making. The aim of this study was to set a research agenda for the PREFER project. The first step involved using the results of literature reviews and stakeholder interviews to develop over 100 questions on the methodology, design, conduct, and application of preference studies. The most important research needs identified were related to four high-level concepts: evidentiary standards, assessment of preference heterogeneity, means to minimize patient burden, and means to maximize patient understanding of concepts presented in preference studies. From this list, questions which could be examined in a patient preference case study and which focused on more promising preference methods which had not been well-studied as of March 2018 (according to PREFER partners, stakeholders, and external scientific advisors) were included in a ranking exercise where respondents were asked to identify their top five questions to be addressed in a PREFER case study. In total, n=33 partners in the PREFER consortium responded to this survey resulting in the identification of 17 prioritized research questions related to three themes: the reliability and validity of preference outcomes, the generalizability and transferability of results, and the impact of educational materials. The research in this thesis presents evidence to help answer questions from this list.

In **Chapters 2 and 3** we looked at how generalizable preferences are from one specific population in a disease to different populations in that or related diseases. In **Chapter 2** we present a study assessing the preferences of patients with diabetes for glucose-monitoring technologies using a discrete choice experiment. Adults with type 1 or 2 diabetes from the Netherlands (n=226) and Poland (n=261) completed an online discrete choice experiment in which they had to choose between hypothetical glucose monitors described using seven

attributes: precision, effort to check, number of finger pricks required, risk of skin irritation, information provided, alarm function and out-of-pocket costs. We assessed the attribute relative importance and calculated expected uptake rates and willingness to pay (WTP) using the outputs of a panel mixed logit analysis. For both countries, the most important attribute was monthly out-of-pocket costs. Beyond this we found large amounts of heterogeneity of preferences not only between countries, but between different patient sub-groups in each country. Polish respondents were more likely than Dutch respondents to choose a glucose-monitoring device over a standard finger prick and had higher WTP for a device (€65.01 vs €27.74). Dutch respondents had higher WTP for device improvements in 'effort to check' (€11.32 vs €3.55) and reducing the number of fingerpricks a device requires (€32.71 vs €13.35). Patients who were younger, had type 1 diabetes, and who currently used a device to monitor blood glucose had higher estimated uptake rates of new glucose-monitoring technologies compared to older patients, patients with type 2 diabetes, or patients who currently only used fingerpricking to monitor blood glucose levels.

In **Chapter 3** we presented a study demonstrating how Q-methodology could be used to identify common opinions about the most important areas of unmet medical needs in patients with two types of neuro-musculoskeletal diseases (myotonic dystrophy type 1 and mitochondrial disorders) and their caregivers. 75.6% of patients and 90% of caregivers said it was 'easy' or 'very easy' to understand the Q-methodology questions. Seven factors representing clinically meaningful viewpoints about unmet medical needs were identified. The most common viewpoint was related to improving physical capability, a viewpoint that was a high priority for patients and caregivers of both disease types. Patients with myotonic dystrophy type 1 and their caregivers tended to focus primarily on muscle strength, energy and endurance, or reducing the side-effects of liver damage associated with treatment. Those with mitochondrial disorders and their caregivers tended to focus more on pain in joints and muscles and improving basic functioning like speech and communication. This study highlighted the feasibility of using Q-methodology to understand patient priorities in a rare disease patient population.

In **Chapter 4** we presented the outcomes of a case study in which respondents were asked to complete a discrete choice experiment which varied in the way that an attribute was framed when presented to participants. The aim of this study was to understand the impact of attribute framing on preferences. In this study, respondents from the general Swedish population (n=1119) were asked about their preferences for antibiotic treatments using five attributes. Four attributes were static (Contribution to Antibiotic Resistance, Treatment Duration, Likelihood of Side-Effects, and Costs), but a fifth treatment attribute regarding treatment effectiveness was framed in three ways: Effectiveness, Failure Rate, or both. We found that attribute framing impacted not only the valuation of the attribute in question but also concurrent valuation of other attributes in a DCE, altering the utility of the alternative. While this did not have an

impact on understanding the relative importance of attributes ('Contribution to Antibiotic Resistance' and 'Costs' were still the most important attributes for all participants regardless of effectiveness framing), it may have an impact on other outcomes such as willingness to pay or maximum acceptable risk for which these differences are sensitive. Thus, when developing a preference study the framing of attributes should be considered not only in regards to the impact on that attribute but also in regards to the impact on other attributes.

In **Chapters 5 and 6** we compared the results of simpler/cheaper methods (Swing Weighting with Direct Rating) to more complex/expensive methods (Discrete Choice Experiments). In **Chapter 5** these methods were compared in a case study assessing patient preferences for glucose monitoring for self-management of diabetes. In this study a sample of Dutch adults with type 1 or 2 diabetes (n=459) completed an online survey assessing their preferences for glucose-monitoring devices, consisting of both a DCE and a Swing Weighting with Direct Rating (SW-DR) exercise. In the SW-DR task, respondents first rank attribute improvements according to their importance and then give points to these improvements to indicate relative importance (on a scale from 0-100). Half the sample completed the DCE first, the other half completed the SW-DR first. The relative importance of the attributes derived from each method were compared. Participants reported the DCE as being easier to understand and answer compared to the SW. Both methods revealed that cost and precision of the device were the most important attributes. However, the difference in relative importance between the two was marked as the most important attribute in the DCE was 14.9 times as important as the least important attribute. In the SW-DR this was difference was only 1.4 times with the relative attribute weights derived from the SW being almost evenly distributed between all attributes. The findings from the SW-DR task were found in both the swing weighting task analyzed using the rank order centroid approach as well as the direct rating task.

Similarly, in **Chapter 6** both DCEs and SW-DR were compared but this time in a case study measuring the treatment preferences of patients with non-small cell lung cancer (n=307) in Italy and Belgium. Most respondents found both tasks very easy or easy to understand and answer. We found that 'Chance of 5-year Survival' and 'Risk of Extreme Tiredness' were the most important attributes in both countries regardless of the method used to assess preferences. Here again, the magnitude in differences between the most important attribute (Chance of 5-year Survival) and other attributes was much larger in the DCE outcomes than in the SW-DR outcomes. In addition, the relative ranking and weight of the less important attributes differed significantly between the DCE and SW-DR. In a small pilot study, we assessed whether the results of the swing weighting task would be different if a point allocation task (dividing 100 points between attributes) was used instead of DR which forces respondents to tradeoff between different attributes like the forced choice of DCEs. We found that the results were more comparable to DCE outcomes but there was now a greater risk of dominant attributes being overvalued.

In **Chapter 7** we conducted a study to understand how preference outcomes may be different when respondents are given information prior to a preference study using a video-based educational tool versus traditional text-based education. In this study, patients with diabetes from the Netherlands (n = 459) and Poland (n = 522) were randomized to receive information about glucose monitoring and the attributes used in the preference assessment in either a text or a video with animations and a voiceover. While some differences in preferences were found between the respondents who saw the different educational materials, no interpretable pattern of differences in the relative importance of attributes could be identified. Examination of the meta-data reporting on the time that respondents spent in different parts of the tools showed that patients spent less time in the educational material than would be necessary to fully review all the content. This was found in both types of educational material in both countries indicating that engagement with educational materials may be a primary issue in preference studies.

In **Chapter 8** we revisited the aim in Chapter 1 and conducted a survey of preference research stakeholders from industry, academia, consultancy, HTA/Regulatory, and patient organizations regarding what methodological topics and questions for future studies they think are important to address to increase acceptance of preference methods and the use of their results by decision makers. This study consisted of a prioritization task, a best-worst scaling case 1 (BWS) instrument, and open-ended questions. In total, n=101 participants responded to the survey invitation with n=66 completing both the prioritization and BWS tasks. The most important research topics related to a mix of applied and methodological research topics including synthesis of preferences across studies, transferability across populations or related diseases, and methods topics including comparison of methods and non-discrete choice experiment methods. Differences in prioritizations were found between respondents whose primary affiliation was academia vs. other stakeholders. Academic researchers prioritized methodological and/or less studied topics, whereas other stakeholders prioritized applied research topics relating to consistency of practice.

GENERAL DISCUSSION AND CONCLUSIONS

As the field of patient preference research grows, empirical evidence addressing methodological questions regarding the assessment and use of this information will continue to be generated and new questions will arise. This thesis presented information regarding prioritized research questions from stakeholders who are interested in using patient preference information, but questions remain. Promising and less utilized methods to assess patient preferences should be tested not only by comparing their outcomes to more trusted methods, but also by testing how they can be applied to decision making and whether decision makers will accept their results. While there will likely never be a "correct" way to frame attributes in a preference study, the impact of framing needs to be assessed when developing a study. Aligning the framing with

other studies and how these attributes are framed in clinical care will at least increase the generalizability of preference research. Understanding whether preference outcomes generated in one treatment context are transferable to others can help to ensure that decision makers are able to use patient values to guide decision making and that research resources are only used when necessary. In this vein, more work needs to be done to understand how different use cases of transferred preference information may have different evidentiary standards. Finding ways to increase engagement with preference assessment materials will help to support the use of patient preferences as being informed and reflective of patient values. Researchers should be cautious when using preference information without assessing whether respondents were engaged with and attended to the relevant aspects of the preference tasks as this can have serious implications on the preference outcomes. A deeper understanding of how preferences change over time and the factors that impact these changes (such as after experiencing treatments or as disease experience develops) would greatly benefit our understanding of the patient experience and better develop medical care that meets their needs, although this is a difficult subject to address. Collaborative efforts between regulators, industry stakeholders, and academic research partners will help not only in answering future methodological questions but also in ensuring that the needs of different stakeholders are being met.

The field of patient preference as applied to medical product development is relatively young meaning the different use cases and methods to assess patient preferences are still being understood. Greater scrutiny of the field will come as the use of patient preference information grows and stakeholders seek to ensure that the information resulting from these studies is of sound methodological quality. More research should be done to ensure that when decision makers use preference information, they can have confidence that they are making decisions with a good understanding of what patients prefer.

Samenvatting in het Nederlands

Het meten van voorkeuren van patiënten (ook bekend als patiëntenpreferentie informatie (PPI)) wordt gedefinieerd als een kwalitatieve of kwantitatieve beoordeling van het relatieve belang of de aanvaardbaarheid van specifieke alternatieven of keuzes tussen uitkomsten (of andere attributen) van verschillende gezondheidsinterventies. In eenvoudigere bewoordingen wordt bedoeld dat er wordt gemeten wat belangrijk of waardevol is voor patiënten. De waarde en het nut van het meten van patiënten preferenties wordt steeds meer erkend, omdat dit past bij de verschuiving naar een meer patiëntgerichte benadering tijdens de ontwikkeling van medische producten. PPI kan bijvoorbeeld worden gebruikt om veelbelovende onderzoeksgebieden te identificeren binnen de farmaceutische industrie of om aan te tonen dat medische producten een acceptabele balans tussen risico's (bijwerkingen) en effectiviteit hebben. PPI kan ook worden gebruikt door toezichthouders om te bepalen of medische producten gewenst zijn door patiënten en aansluiten op de behoeften van patiënten. Daarnaast kan het worden gebruikt door patiëntvertegenwoordigers en wetenschappelijke onderzoeksorganisaties om onderzoeksgebieden te identificeren die meer aandacht zouden moeten krijgen. Dit zijn slechts enkele voorbeelden, maar in alle gevallen willen belanghebbenden die deze informatie gebruiken bij beleidsbeslissingen zeker zijn dat de gebruikte PPI van goede methodologische kwaliteit is. Dit proefschrift heeft als doel empirisch bewijs te leveren met betrekking tot methodologische vraagstukken rondom het meten van PPI.

In **Hoofdstuk 1** worden de uitkomsten van een onderzoek beschreven waarbij er een enquête gestuurd is naar leden van het PREFER-consortium om methodologische vragen te identificeren en te prioriteren met betrekking tot het uitvoeren van PPI onderzoek dat als doel heeft besluitvorming rondom medische producten te ondersteunen. Het doel van deze studie was het opstellen van een onderzoeksagenda voor het IMI-PREFER project, waarin de rol van PPI in besluitvorming langs de medicijn-ontwikkelketen centraal stond. Op basis van literatuuronderzoek en interviews met belanghebbenden is er eerst een lijst met meer dan 100 onderzoeksvragen ontwikkeld over de methodologie, het ontwerp, de uitvoering en de toepassing van onderzoek naar patiëntenpreferenties. De belangrijkste onderzoeksvragen die zijn geïdentificeerd konden worden onderverdeeld in vier groepen: bewijsstandaarden, beoordeling van heterogeniteit in preferenties, methoden om de belasting op de patiënt te minimaliseren, en methoden om het begrip van de belangrijkste onderdelen in PPI onderzoek voor patiënten te maximaliseren. Van deze lijst werden 27 vragen, die empirisch konden worden onderzocht in een case studie en die gericht waren op veelbelovende methoden voor onderzoek naar patiënten preferenties, opgenomen in een prioriteringsoefening. In deze oefening werd respondenten gevraagd om hun vijf belangrijkste vragen te identificeren die in een PREFER-case studie onderzocht zouden moeten worden. In totaal hebben n=33 partners in het PREFER-consortium gereageerd op deze enquête, resulterend in de identificatie van 17 geprioriteerde onderzoeksvragen gerelateerd aan drie thema's: 1) de betrouwbaarheid en validiteit van de uitkomsten van patiënten preferentie onderzoek, 2) de generaliseerbaarheid en de vertaalbaarheid van resultaten uit eerder onderzoek naar andere contexten, en 3) de impact

van educatie materiaal in vragenlijsten. In dit proefschrift worden onderzoeksvragen uit deze lijst door middel van empirisch onderzoek beantwoord.

In de **Hoofdstukken 2 en 3** is gekeken naar hoe generaliseerbaar de patiëntenpreferenties zijn van een specifieke populatie met een specifieke ziekte naar andere populaties met dezelfde of een andere (gerelateerde) ziekte. In **Hoofdstuk 2** wordt een studie gepresenteerd waarin de preferenties van patiënten met diabetes voor glucosemonitoringtechnologieën werden gemeten met behulp van een discrete keuze-experiment (DCE). Volwassenen met diabetes type 1 of 2 uit Nederland (n=226) en Polen (n=261) hebben een online vragenlijst ingevuld met diverse scenario's waarin ze werd gevraagd om te kiezen tussen hypothetische glucosemeters beschreven aan de hand van zeven attributen: precisie, mate van inspanning om glucosewaarden te controleren, aantal vingerprikken per dag, risico op huidirritatie, het type informatie dat beschikbaar is, alarmfunctie en 'out-of-pocket' kosten. Er zijn verschillende uitkomsten gerapporteerd op basis van de resultaten van een panel-mixed logit analyse, namelijk het relatieve belang van elk attribuut, de te verwachten gebruikspercentages en de betalingsbereidheid (ook bekend als willingness-to-pay; WTP). Voor beide landen bleek de maandelijkse 'out-of-pocket' kosten het belangrijkste attribuut in de keuze van patiënten om een glucosemeter te gebruiken. Daarnaast is er heterogeniteit in de preferenties van patiënten geïdentificeerd, niet alleen tussen landen, maar ook tussen verschillende subgroepen van patiënten binnen elk land. Poolse respondenten kozen vaker voor een glucosemeter in plaats van de standaard vingerprik en hadden een hogere WTP voor een apparaat (€ 65,01 versus € 27,74) dan Nederlandse respondenten. Nederlandse respondenten hadden een hogere WTP om de 'inspanning om te controleren' te verminderen (€ 11,32 versus € 3,55) en het aantal dagelijkse vingerprikken dat nodig is om glucosewaarden te controleren te verminderen (€32,71 versus € 13,35). Patiënten die jonger waren, diabetes type 1 hadden, en die al bekend waren met het gebruik van apparatenom de bloedglucose te controleren waren meer bereid om nieuwe technologieën voor glucosemonitoring te gebruiken in vergelijking met oudere patiënten, patiënten met diabetes type 2, of patiënten die alleen vingerprikken gebruikten om bloedglucose te meten.

Hoofdstuk 3 presenteert een studie die aantoonde hoe Q-methodologie kan worden gebruikt om gemeenschappelijke meningen te identificeren over de belangrijkste medische behoeften van patiënten met twee soorten neuro-musculoskeletale aandoeningen (myotone dystrofie type 1 en mitochondriale aandoeningen) en hun verzorgers. Voor deze ziekten zijn weinig medicijnen beschikbaar, het onderzoek was dan ook gericht op het identificeren van problemen waar een toekomstig medicijn aan zou kunnen bijdragen. In dit onderzoek werd een online vragenlijst uitgezet onder patiënten en hun verzorgers in 5 landen op drie continenten. Deelnemers moesten aangeven op welke beperkingen van de ziekte het denkbeeldige medicijn zich zou moeten richten, maar ook welke bijwerkingen van zo'n medicijn het meest ongewenst waren. Er werden zeven factoren geïdentificeerd die klinisch betekenisvolle standpunten vertegenwoordigen over

medische behoeften van deze patiënten. Het meest voorkomende standpunt was gerelateerd aan het verbeteren van fysieke mogelijkheden, deze mening werd gedeeld door patiënten en zorgverleners van beide ziektetypes. Patiënten met myotone dystrofie type 1 en hun verzorgers focusten voornamelijk op spierkracht, energie en uithoudingsvermogen, of het verminderen van de bijwerkingen 'leverschade'. Patiënten met mitochondriale aandoeningen en hun verzorgers vonden pijn in gewrichten en spieren en het verbeteren van basisfuncties zoals spraak en communicatie juist belangrijker. In totaal gaf 75,6% van de patiënten en 90% van de zorgverleners aan dat het 'gemakkelijk' of 'zeer gemakkelijk' was om de Q-methodologievragen te begrijpen Deze studie benadrukt de haalbaarheid van het gebruik van de Q-methodologie om de prioriteiten van een patiëntenpopulatie met een zeldzame ziekte te identificeren.

In **Hoofdstuk 4** worden de resultaten van een case studie gepresenteerd waarin respondenten werd gevraagd een vragenlijst met een DCE in te vullen waarbij de bewoording van een specifiek attribuut op verschillende wijzen (in positieve en negatieve bewoording) werd gepresenteerd aan deelnemers. Het doel van deze studie was om de impact van attribuut framing op preferenties te onderzoeken. In deze studie werden respondenten uit de Zweedse bevolking (n=1119) gevraagd naar hun preferenties voor antibioticabehandelingen aan de hand van vijf attributen. Vier attributen waren statisch (bijdrage aan antibioticaresistentie, duur van de behandeling, risico op bijwerkingen en kosten), maar een vijfde attribuut met betrekking tot de effectiviteit van de behandeling werd op drie manieren verwoord: kans dat de behandeling effectief is, kans dat de behandeling niet werkt, of beide. Attribuut framing had niet alleen invloed op de waardering van het betreffende attribuut, maar ook op gelijktijdige waardering van andere attributen in een DCE. Hoewel de framing geen invloed had op het relatieve belang van de attributen ('Bijdrage aan antibioticaresistentie' en 'Kosten' waren nog steeds de belangrijkste attributen voor alle deelnemers ongeacht de framing van effectiviteit) kan het wel impact hebben op andere uitkomsten (zoals betalingsbereidheid of maximaal aanvaardbaar risico). Bij het ontwikkelen van een preferentieonderzoek is het belangrijk om aandacht te besteden aan de framing van attributen, niet alleen vanwege de impact van framing op het attribuut zelf, maar ook omdat uitkomsten van andere attributen daardoor beïnvloed worden.

In **Hoofdstukken 5 en 6** zijn de resultaten van eenvoudigere/goedkopere methode (Swing Weighting met Direct Rating) vergeleken met complexere/duurdere methode (DCEs) voor het meten van patiëntpreferenties. In **Hoofdstuk 5** werden deze methoden vergeleken in een case studie waarin de preferenties van patiënten voor glucosemonitoring voor zelfmanagement van diabetes werden onderzocht. In deze studie vulde een steekproef van Nederlandse volwassenen met diabetes type 1 of 2 (n=459) een online enquête in om hun preferenties voor diverse vormen van glucosemeting te geven. De enquête bestond uit zowel een DCE als een Swing Weighting met Direct Rating (SW-DR). In de SW-DR werd aan respondenten gevraagd eerst de verbetering van zeven attributen te rangschikken op basis van belangrijkheid. Vervolgens mochten respondenten punten aan deze verbeteringen toekennen (op een schaal van 0-100).

De helft van de steekproef voltooide eerst de DCE, de andere helft voltooide eerst de SW-DR. Het relatieve belang van de attributen afgeleid uit beide methode werd vergeleken. Deelnemers gaven aan dat de DCE gemakkelijker te begrijpen en te beantwoorden was in vergelijking met de SW. Beide methoden toonden aan dat de kosten (eigen bijdragen) en precisie van het apparaat relatief de belangrijkste attributen waren. Het belangrijkste attribuut was in de DCE echter 14,9 keer zo belangrijk ten opzichte van het minst belangijke attribuut, terwijl dit voor de SW-DR 1,4 was. De relatieve attribuutgewichten afgeleid van de SW waren bijna gelijk verdeeld over alle attributen, dit was onafhankelijk van de gebruikte statistische toets om het relatieve belang te berekenen. Uit dit onderzoek is gebleken dat preferentieuitkomsten bij metingen met deze twee methoden verschillend kunnen zijn als het gaat om het relatieve belang van de attributen.

Ook in **Hoofdstuk 6** werd een DCE met SW-DR vergeleken, maar dit keer in een case studie waarin de behandelvoorkeuren van patiënten met niet-kleincellig longkanker (n=307) in Italië en België werden gemeten. De 'Kans op 5-jaarsoverleving' en 'Risico op extreme vermoeidheid' waren de belangrijkste attributen in beide landen, ongeacht de methode die werd gebruikt om preferenties te meten. Ook hier was de grootte van de verschillen tussen het belangrijkste attribuut ('Kans op 5-jaarsoverleving') en andere attributen veel groter op basis van de DCE-uitkomsten dan op basis van de SW-DR-uitkomsten. Bovendien verschilde het relatieve belang en het relatieve gewicht van de minder belangrijke attributen significant tussen de DCE en SW-DR. De meeste respondenten vonden beide taken (erg) gemakkelijk te begrijpen en te beantwoorden. In een kleine pilotstudie is beoordeeld of de resultaten van de SW-taak anders zouden zijn als een puntentoewijzingstaak (het opsplitsen van 100 punten tussen attributen) zou worden gebruikt in plaats van DR. Hierdoor worden respondenten gedwongen een afweging te maken tussen verschillende attributen, zoals ook bij een DCE. De resultaten van de SW in deze pilotstudie waren meer vergelijkbaar met DCE-uitkomsten. Echter was er nu een groter risico dat dominante attributen overgewaardeerd werden.

In **Hoofdstuk 7** is onderzoek uitgevoerd om te begrijpen hoe uitkomsten van een patiënten preferentie onderzoek kunnen verschillen wanneer respondenten voorafgaand aan de vragenlijst op verschillende manieren (traditioneel met tekst of op basis van een video) uitleg krijgen over de inhoud van het keuze experiment. In deze studie werden patiënten met diabetes uit Nederland (n = 459) en Polen (n = 522) gerandomiseerd om tekstueel of via een video met animaties en voice-over informatie te ontvangen over glucosemonitoring en de gebruikte attributen (en levels) van de DCE. Hoewel er enkele verschillen in preferenties werden gevonden tussen de respondenten die de verschillende vormen van het educatieve materiaal zagen, kon er geen inzichtelijk patroon van verschillen in het relatieve belang van attributen worden vastgesteld. Onderzoek naar de tijd die respondenten aan verschillende delen van het educatieve materiaal besteedden (gemiddeld 2.36 tot 3.84 minuten), toonde aan dat patiënten minder tijd aan het educatieve materiaal besteedden dan nodig zou zijn om alle inhoud volledig

te bekijken. Dit werd gevonden voor beide soorten educatief materiaal en in beide landen, wat aangeeft dat het verbeteren van de betrokkenheid van deelnemers bij het educatieve materiaal van een DCE van groot belang is in toekomstig onderzoek naar patiënten preferenties

In **Hoofdstuk 8** zijn de uitkomsten van Hoofdstuk 1 opnieuw bekeken. Vier jaar na de eerste studie naar de belangrijkste te beantwoorden onderzoeksvragen over methoden om patiëntpreferenties te meten werd een nieuwe enquête gehouden onder belanghebbenden op het gebied van patiënten preferentie onderzoek. Dit keer werden niet alleen betrokkenen bij IMI-PREFER gevraagd, er was een breed scala aan deelnemers uit de farmaceutische industrie, de academische wereld, adviesbureaus, HTA/regelgevende instanties en patiëntenorganisaties. Hen is gevraagd wat de belangrijkste methodologische onderwerpen en vragen zijn voor toekomstig onderzoek dat bij zou moeten dragen aan de acceptatie van methoden voor patiënten preferentieonderzoek en het gebruik van de resultaten door beleidsmakers. Dit onderzoek bestond uit een prioriteringstaak, een best-worst scaling case 1 (BWS)-instrument en open vragen. In totaal reageerden 101 deelnemers op de enquête en 66 deelnemers voltooiden zowel de prioriteringstaak als de BWS-taak. De belangrijkste onderzoeksthema's waren synthese van uitkomsten van verschillende patiënten preferentieonderzoeken, de vertaalbaarheid van uitkomsten van preferentieonderzoek naar andere contexten, zoals andere landen en andere patiëntenpopulaties, en vragen over methoden, waaronder het vergelijken van verschillende methoden voor het meten van preferenties en onderzoek naar andere methoden dan DCE. Er werden verschillen in prioritering gevonden tussen respondenten uit de academische wereld versus andere belanghebbenden. Academische onderzoekers gaven prioriteit aan methodologische en/of minder bestudeerde onderwerpen, terwijl andere belanghebbenden prioriteit gaven aan toegepaste onderzoeksthema's, zoals de keuze van methoden.

ALGEMENE DISCUSSIE EN CONCLUSIES

Naarmate het onderzoeksveld rondom patiëntpreferenties groeit, zal er steeds meer empirisch bewijsmateriaal worden gegenereerd dat methodologische vragen beantwoord, maar er zullen ook nieuwe vragen ontstaan. Dit proefschrift presenteerde informatie met betrekking tot geprioriteerde onderzoeksvragen van belanghebbenden die geïnteresseerd zijn in het gebruik van PPI. Er zijn echter ook nieuwe aandachtsgebieden voor toekomstig onderzoek geïdentificeerd. Veelbelovende en minder gebruikte methoden om de preferenties van patiënten te meten moeten worden getest door hun resultaten te vergelijken met meer vertrouwde methoden. Daarnaast zal ook moeten worden bekeken of de resultaten van deze methoden kunnen worden toegepast op besluitvormingsprocessen voor medische producten en of besluitvormers de resultaten van dergelijk onderzoek zullen accepteren. Hoewel er waarschijnlijk nooit één 'juiste' manier zal zijn om attributen in een preferentieonderzoek te verwoorden, moet de impact van het 'framen' van attributen worden beoordeeld tijdens het ontwikkelen van instrumenten waarmee patiëntpreferenties gemeten worden. Het afstemmen van de framing op andere

onderzoeken en op de framing van deze attributen in de klinische zorg zal op zijn minst de generaliseerbaarheid van preferentieonderzoek vergroten. Inzicht in de vraag of de resultaten van patiëntpreferentieonderzoek die in een bepaalde behandelcontext worden gegenereerd, overdraagbaar zijn naar andere contexten en populaties, kan ervoor zorgen dat beleidsmakers patiëntpreferenties kunnen gebruiken als leidraad voor besluitvorming zonder dat ze elke keer een nieuwe studie moeten uitvoeren en dat onderzoeksmiddelen alleen worden gebruikt als dat nodig is. Gelijktijdig is er meer onderzoek nodig naar verschillende bewijsnormen voor het toepassen van PPI op andere populaties en contexten.. Het identificeren en testen van manieren om de betrokkenheid van deelnemers bij het educatieve materiaal in keuze experimenten (en bij het onderzoek zelf) te vergroten, zal helpen om de beslissingen van patiënten in preferentieonderzoek te kunnen bestempelen als goed geïnformeerde keuzes die een afspiegeling zijn van preferenties van patiënten. Onderzoekers moeten voorzichtig zijn bij het gebruik van preferentie informatie wanneer het onduidelijk is of respondenten betrokken waren bij en aandacht besteed hebben aan de educatieve materialen van het keuze experiment, aangezien dit van invloed kan zijn op de uitkomsten van het onderzoek. Hoewel dit een moeilijk onderwerp is om te onderzoeken, zou een beter begrip van hoe preferenties in de loop van de tijd veranderen en de factoren die van invloed zijn op deze veranderingen (zoals na het ervaren van behandelingen of naarmate de ziekte zich ontwikkelt) het begrip van de preferenties van patiënten enorm ten goede komen. Dit kan er vervolgens toe leiden dat medische zorg beter kan inspelen op de behoeftes van patiënten.

Het onderzoek naar patiënten preferenties binnen het kader van beslissingen over de ontwikkeling van medische producten is relatief jong. Dit betekent dat de verschillende 'use cases' en methoden om de preferenties van patiënten te meten nog steeds worden ontwikkeld en onderzocht. Er zal meer aandacht voor het veld komen naarmate het gebruik van informatie over patiëntenpreferenties groeit en belanghebbenden ervoor willen zorgen dat de informatie die uit deze onderzoeken voortkomt van degelijke methodologische kwaliteit is. Verder onderzoek op het gebied van het meten van preferenties kan er aan bijdragen dat beleidsmakers patiëntpreferenties meewegen bij hun beslissingen, en dat zij vertrouwen dat er op de juiste manier gemeten is wat patiënten belangrijk vinden in hun zorg. Gezamenlijke inspanningen tussen regelgevers, belanghebbenden uit de sector, en academische onderzoekspartners zullen niet alleen helpen bij het beantwoorden van toekomstige methodologische vragen, maar ook om ervoor te zorgen dat aan de behoeften van verschillende belanghebbenden wordt voldaan.

Acknowledgments

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

- from “Little Gidding” in *Four quartets* by T.S. Eliot

The squiggle on the cover of this book was created by Damien Newman as a way to visualize the process of design. When we hold a finished product in our hand, we often don't have a good understanding of everything that went into getting to this point. The messy periods of learning and uncertainty, searching for ideas that are relevant and should be pursued. Testing these ideas and returning again to the learning period to look for new directions. Finally finding something worth pursuing; testing, trying, and refining it over and over again until the final product becomes clear and can be developed and sent out into the world. The time and energy that goes into this process is often overlooked and underappreciated when all you have in your hands is the final product.

Much like the process of design, the process of research and writing a thesis works the same way. The initial messy learning period discovering the foundations of the field while seeing how new research is pushing the field forward, looking for insights from this knowledge, developing questions and hypotheses worth testing, rejecting these hypotheses and starting over, encountering setbacks, starting over, testing, analyzing, testing again, realizing you left a variable out of the syntax for an analysis that takes an entire weekend to run and putting it in before testing again, finally getting to a point where you are confident in your results so you can write up what you learned and communicate this to others, and at the end is a manuscript. This simple book is the result of years of work, time and energy spent getting to this point. It is a hard process to follow and if you're lucky you don't do this alone. I am lucky.

Throughout my PhD I have been supported by many people, but the two people that supported me the most (from before I was even officially a PhD student) are Ardine de Wit and Jorien Veldwijk.

To Ardine, this thesis was only possible through your patience and support. Whether this was by supporting me in pursuing ambitious ideas that didn't fit the mold of other case studies and didn't actually pan out (e.g., eye-tracking), going on long walks with me during the pandemic when everything was falling apart and nothing was progressing, or pushing me to just get this thesis finished. Every step I have taken to get here has been possible because of the patience and support you gave me. When I would talk with other PhDs about the support you offered, they would often respond with incredulity that a promotor would offer up so much of their time and energy to support a PhD, but you did. You have a positive impact on the world and those around you and I hope you know this. Yours is the example I want to follow.

To Jorien, thank you for answering all the "quick" questions that I always managed to think of as we were wrapping up meetings even when these questions required hours of back and forth and multiple day correspondences before I understood the answer. I still can't believe that you didn't snap when you were helping me understand scale parameters. It was through your course on discrete choice experiments that the concept of patient preferences became real to me. I hope that someday I have the same impact on curious minds that you regularly have.

To Rachael DiSantostefano, you handed me responsibility for writing the 'stick-in-my-eye paper' early in my PhD and pulled me through that painful writing process in a way that was not only bearable, it was downright enjoyable! If there was one moment that I can point at and say "this was when I knew I could do this," it was getting that paper published. Your constant laughter and support through all our work since has been a joy to me. I try to emulate your example of just getting something down on paper even if it is just "Blah, Blah, Blah, Blah, Blah- come back to this later." Thank you!

To Karin Bywall, our calls to talk about the "joys" of having children and a family while doing PhD research kept me (almost) sane. Thank you for all the hours you spent listening to me and sharing your own stories so I knew I wasn't the only one. I hope we can continue our digital fikas for many years to come.

To my coauthors, Esther de Bekker-Grob, Bennett Levitan, Chiara L Whichello, Ellen Janssen, Karin Groothuis-Oudshoorn, Maureen Rutten-van Mölken, Conny Berlin, Mirko Ancillotti, Cecilia Jimenez-Moreno, Eline van Overbeeke, Cathy Anne Pinto, Grainne Gorman, Marie Falahee, Serena Oliveri, Rosanne Janssens, Dario Monzani, Meredith Smith, Luca Bailo, and Rimke Vos, thank you for waiting for me to get over my writer's block and helping to form the content of this thesis. Without you I wouldn't have made it past chapter 2!

To Conny Berlin and Mats Hansson and the other members of the PREFER consortium, I didn't realize in the moment how fortunate I was to be connected with you all. Working together with academic, regulatory, industry, and patient stakeholders towards a shared goal was a unique opportunity that too few people have experienced. The success of PREFER was because of you and I am grateful that I could be a part of that process. Thank you for welcoming me from day 1!

To Megan and Rebecca (the Americans from the Julius Center), whether it was drinking a bit of desk whiskey while kvetching about what we were going through in work and life, or going for long walks to also kvetch about what we were going through in work and life, thank you for the pleasure of your company and the shoulder to lean on when things weren't going so great. I look forward to many more long walks and 5 course vegan meals together.

To Hugo, you were the first person I thought of when asked who I wanted to be my paranimf and support me during my defense. I have valued your friendship, honesty, and lack of filter since the first day we met. Thank you for being my friend.

To my Mom, you have tirelessly supported me my entire life. You have been there through all the highs and the many, many lows and never once waivered in your support even when it cost you so much. From the conversations with school administrators when I was almost expelled from grade school, to the hours spent driving around to look at colleges, to flying out to Hawaii to make upside-down pineapple cake with me and Viv, to popping into the Netherlands on

your trips to and from helping people in the last mile in Africa, I always knew that you would be there for me whenever I need it.

To my Dad, I think it's pretty telling that the PhD in the family is a Prep graduate. Just saying. When I was getting my master's degree in psychology, I was looking at different ways to measure patients' perceptions of being involved in decisions around their medical care. I remember talking to you about it before realizing that there, at the number 4 author spot, was one DG Smith. We can both attest that I've never listened to you a day in my life, but somehow your example of caring for the patient as a person with valuable insights that should be considered in medical decision-making got into my head. Thank you for setting this example (even if subconsciously). Your humor even when discussing tough subjects and the house full of music are examples I live by today. This thesis follows on the work that you have done both academically and as a father.

To my sisters Maurie and Cara, when you're not located close to family it can be easy to drift away. You never let that happen even when I wasn't the best at responding. The love you share only makes these bonds stronger. You're a big part of the reason Finnley and Owen always ask to go back to Philadelphia (even if Philly does smell like farts while the Netherlands smells like strawberries). To my sister Julie, it's safe to say that the only reason I'm here right now is because I went to visit you in New Zealand and met Viv in that hostel in Dunedin. Nobody saw this being a possible result of that trip! Never stop being you.

To all of my supportive family including my Aunt Bobbi and stepmom Ann, thank you for your unwavering support for all these years. From sparking a joy in roller coasters to making sure that I was welcome in Vermont, I have never once doubted that I had people who loved and cared for me.

To my Dutch family, José, Dennis, Nadja, and Ralf, thank you for letting me escape to your house for intense writing weekends and for camping by us to help Viv with the kids so that I could work nights and weekends to get this done. Sometimes you just need to push through to get things done and because of you I could do that.

To Finnley and Owen, I did this in spite of and because of you both. I have never loved anything as much as I love you both. The countless nights of sleep deprivation, interrupted thought processes, and guest appearances during meetings where you had to waive to everyone (even when it was during a conference presentation) didn't help me finish this thesis, but the joy and laughter you bring to my life are the motivation to "nooit opgeven". Now please, just 'go to bed'/'put your clothes in the laundry basket'/ 'brush your teeth'/ 'put your shoes on'/ 'listen'/ 'let me close my eyes for five minutes'/'do whatever else I'm asking you to do' and no, you can't have any snoepjes (until when mom isn't looking).

To Vivian, you are the one person who has truly been there with me every step of the way. From the moment when I first said I wanted to get out of the Air Force and go back to university to now, you have been the constant. You have seen me sitting at my laptop still typing as the sun rises, been there through the successes and the tears of frustration, pushed me when I needed to be pushed, and given me the space when I needed it to pursue my academic ambitions (even if that space was cut short by snoring German campers). You bore the burden of all the weekends and vacations I missed or spent distracted while working on my research. Without you none of this would have been possible and for that I will always be thankful. Your eeuwige student is finally finished with his studies and can now "get a real job."

Publications and Presentations

PUBLICATIONS AS PART OF THIS THESIS

Smith, I.P., Disantostefano, R.L., De Bekker-Grob, E.W., Levitan, B., Berlin, C., Veldwijk, J., De Wit, G.A., 2021. Methodological Priorities for Patient Preferences Research: Stakeholder Input to the PREFER Public–Private Project. The Patient - Patient-Centered Outcomes Research 14, 449–453.. doi:10.1007/s40271-021-00502-6

Smith, I.P., Ancillotti, M., de Bekker-Grob, E.W. and Veldwijk, J., 2022. Does It Matter How You Ask? Assessing the Impact of Failure or Effectiveness Framing on Preferences for Antibiotic Treatments in a Discrete Choice Experiment. Patient preference and adherence, pp.2921-2936.

Smith, I.P., Whichello, C.L., Veldwijk, J., Rutten-van Mölken, M.P., Groothuis-Oudshoorn, C.G., Vos, R.C., de Bekker-Grob, E.W. and De Wit, G.A., 2023. Diabetes patient preferences for glucose-monitoring technologies: results from a discrete choice experiment in Poland and the Netherlands. BMJ Open Diabetes Research and Care, 11(1), p.e003025.

Smith, I.P., Whichello, C.L., de Bekker-Grob, E.W., Mölken, M.P.R.V., Veldwijk, J. and de Wit, G.A., 2023. The Impact of Video-Based Educational Materials with Voiceovers on Preferences for Glucose Monitoring Technology in Patients with Diabetes: A Randomised Study. The Patient-Patient-Centered Outcomes Research, pp.1-15.

Whichello, C., Smith, I., Veldwijk, J., de Wit, G.A., Rutten-van Molken, M.P. and de Bekker-Grob, E.W., 2023. Discrete choice experiment versus swing-weighting: A head-to-head comparison of diabetic patient preferences for glucose-monitoring devices. PLoS One, 18(7), p.e0283926.

DiSantostefano, R.L., Smith, I.P., Falahee, M., Jiménez-Moreno, A.C., Oliveri, S., Veldwijk, J., de Wit, G.A., Janssen, E.M., Berlin, C. and Groothuis-Oudshoorn, C.G., 2023. Research Priorities to Increase Confidence in and Acceptance of Health Preference Research: What Questions Should be Prioritized Now?. The Patient-Patient-Centered Outcomes Research, pp.1-12.

Veldwijk J, Smith IP, Oliveri S, et al. Comparing Discrete Choice Experiment with Swing Weighting to Estimate Attribute Relative Importance: A Case Study in Lung Cancer Patient Preferences. Medical Decision Making. 2024;0(0). doi:10.1177/0272989X231222421

Other Publications

Oliveri, S., Lanzoni, L., Veldwijk, J., De Wit, G.A., Petrocchi, S., Janssens, R., Schoefs, E., Smith, M.Y., Smith, I., Nackaerts, K. and Vandevelde, M., 2023. Balancing benefits and risks in lung cancer therapies: patient preferences for lung cancer treatment alternatives. Frontiers in Psychology, 14, p.1062830.

Oliveri, S., Lanzoni, L., Petrocchi, S., Janssens, R., Schoefs, E., Huys, I., Smith, M. Y., Smith, I. P., Veldwijk, J., G Ardine, d. W., & Pravettoni, G. (2021). Opportunities and challenges of web-based and remotely administered surveys for patient preference studies in a vulnerable population. Patient Preference and Adherence, 15, 2509-2517. doi:10.2147/PPA.S327006

Monzani D, Petrocchi S, Oliveri S, Veldwijk J, Janssens R, Bailo L, Smith MY, Smith I, Schoefs E, Nackaerts K, Vandevelde M, Louis E, Decaluwé H, De Leyn P, Declerck H, Katz EG, Petrella F, Casiraghi M, Durosini I, Galli G, Garassino MC, de Wit GA, Pravettoni G, Huys I. Patient Preferences for Lung Cancer Treatments: A Study Protocol for a Preference Survey Using Discrete Choice Experiment and Swing Weighting. Front Med (Lausanne). 2021 Aug 2;8:689114. doi: 10.3389/fmed.2021.689114. PMID: 34409049; PMCID: PMC8365300.

Jimenez-Moreno, A.C., van Overbeeke, E., Pinto, C.A., Smith, I., Sharpe, J., Ormrod, J., Whichello, C., de Bekker-Grob, E.W., Bullok, K., Levitan, B. and Huys, I., 2021. Patient preferences in rare diseases: a qualitative study in neuromuscular disorders to inform a quantitative preference study. The Patient-Patient-Centered Outcomes Research, 14, pp.601-612.

Durosini, I., Janssens, R., Arnou, R., Veldwijk, J., Smith, M.Y., Monzani, D., Smith, I., Galli, G., Garassino, M., Katz, E.G. and Bailo, L., 2021. Patient preferences for lung cancer treatment: a qualitative study protocol among advanced lung cancer patients. Frontiers in public health, 9, p.622154.

Jimenez-Moreno, A.C., Pinto, C.A., Levitan, B., Whichello, C., Dyer, C., Van Overbeeke, E., de Bekker-Grob, E., Smith, I., Huys, I., Johansson, J.V. and Adcock, K., 2020. A study protocol for quantifying patient preferences in neuromuscular disorders: a case study of the IMI PREFER Project. Wellcome Open Research, 5.

TECHNICAL RESEARCH REPORTS

2021 IMI-PREFER Use of Educational Materials in Preference studies:Final report describing the process and results of a case study identifying Diabetes patient preferences for glucose monitoring devices.

2021 IMI-PREFER Lung Cancer Case Study Final Report: Final report describing the process and results of a case study identifying Non-Small Cell Lung Cancer patient preferences for immunotherapy.

2021 IMI-PREFER NMD Case Study Final Report: Final report describing the process and results of a case study identifying the unmet medical needs of patients with Neuromuscular disorders.

2019 Methodological Priorities for IMI-PREFER Case studies: Compiled and described over 100 prioritized research questions from IMI-PREFER stakeholders to guide academic and industry case studies.

PRESENTATIONS

2022 ISPOR Europe, Vienna, Austria

Poster Presentation: Do We Really Need a Study Looking at That? A Best-Worst Scaling study assessing methodological research priorities according to the patient preference research community

Poster Presentation: Does the format of educational materials impact diabetes patient preferences for glucose monitoring technology? A randomized study in two countries

2020 ISPOR Europe, Virtual

Poster Presentation: The impact of framing an attribute as failure or effectiveness on preferences for antibiotic treatment in a discrete choice experiment

2020 Lowlands Health Economic Study Group, Virtual

Discussion of outcomes: The impact of framing an attribute as failure or effectiveness on preferences for antibiotic treatment in a discrete choice experiment

About the author

Ian P. Smith was born in New Haven, CT, USA in 1981 and spent most of his childhood in Philadelphia, PA, USA. Ian graduated from St. Joseph's Preparatory High School and, after an unsuccessful attempt of university studies at Earlham College, joined the U.S. Air Force. Ian spent 7 years as an aircraft mechanic in the U.S. Air Force traveling the world fixing planes, but at the end of his contract decided that it was time for something new. After moving to the Netherlands with his partner, he returned to academia and obtained a Bachelor of Sciences degree in Psychology from The Pennsylvania State University in 2013. Early insight into the ways that cognitive processes impact mental health led him to further pursue psychology, obtaining a Master of Science degree in clinical and health psychology from Leiden University in 2015. It was during these studies that he was first introduced to the concept of medical psychology and conducted research looking at how the quality of life of patients with ischemic heart disease is impacted by the presence of comorbid medical conditions. During his internship at the Amsterdam Medical Center, Ian saw how scientific research into the patient experience was used by clinical teams to improve patient care, a cornerstone in his career development. After completing his master's degree at Leiden University, Ian continued in research, first by assessing shared decision making in the clinical context, then by helping to develop a self-management mobile application for diabetics patients at the Leiden University Medical Center. During this research Ian was surprised at how the patient voice was often not heard, either in clinical care or in medical product development. When a PhD position looking at ways to incorporate the patient voice into the medical product development lifecycle opened in the IMI-PREFER project, Ian jumped at the chance to conduct his doctoral research under the supervision of Prof. dr Ardine de Wit and Dr. Jorien Veldwijk at the University Medical Center Utrecht.

METHODOLOGICAL QUESTIONS REGARDING THE ELICITATION OF PATIENT PREFERENCES AND EMPIRICAL EVIDENCE TO HELP ADDRESS THEM

Patient preferences are qualitative or quantitative assessments of the relative desirability or acceptability to patients of specified alternatives or choices among outcomes or other attributes that differ among alternative health interventions. In simpler terms, they are assessments of what patients want or would accept in their care. This PhD thesis presents methodological questions regarding the assessment of patient preferences along with research addressing these questions. The aim of this thesis is to help stakeholders who would potentially use this information to better understand and conduct patient preference studies and support the use of patient preferences in medical product decision-making.

Ian Smith conducted his research at the University Medical Center Utrecht under the supervision of Prof. dr. Ardine de Wit and Dr. Jorien Veldwijk as a part of the IMI-PREFER project. His educational background is in clinical and health psychology, aviation maintenance technology, and epidemiology. He also received a certificate of achievement in weather forecasting from the Pennsylvania State University but doesn't like to brag about that. The focus of his work is trying to find ways to better include the patient voice in healthcare decision-making and promote patient-centered care, whether that is in the consultation room or in medical product development.