Patient (2022) 15:729–738
https://doi.org/10.1007/s40271-022-00601-y

13th Meeting of the International Academy of Health Preference Research

ABSTRACTS

13th Meeting International Academy of Health Preference Research

13TH MEETING OF THE INTERNATIONAL ACADEMY OF HEALTH PREFERENCE RESEARCH

Esther de Bekker-Grob, Erasmus School of Health Policy, Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

Axel Mühlbacher, Health Economics and Management, Hochschule Neubrandenburg, Neubrandenburg, Germany

Established on 15 April 2014, the International Academy of Health Preference Research (IAHPR) is a member-driven, inter-generational organization that promotes educational activities and research with respect to health and health-related preferences. Our aim is to improve decisions about health and healthcare throughout the world by developing, promoting, and supporting health preference research with the widest possible applicability. The 13th Meeting was held on Monday and Tuesday, 5–6 September 2022, and chaired by Michał Jakubczyk and Jorien Veldwijk.

Each of the 17 presentations (abstracts below) had 30 min (20 min for slides, 10 min for discussion) and are listed in order of presentation. The abstract submission system closed on 15 June 2022, and these 17 abstracts were invited out of 34 submissions based on the ratings of the tenured members. The presentations were not streamed online. If willing, presenters may agree to record and distribute recordings of these presentations.

In addition to the podium presentations, the following 3 researchers gave elevator talks (5 minutes for slides; 5 minutes for discussion) to introduce themselves and invite collaboration on their ongoing research (listed in order of their talks): Ann-Kathrin Fischer, Christoph Paul Klapproth and Carina Oedingen. In parallel with the oral presentations, posters were exhibited. Like the abstracts, elevator talks, and poster applications were selected based on the ratings of the tenured members.

Disclaimer

IAHPR, in general, requests that a high standard of science is followed concerning publications and presentations at all its workshops, symposia, and meetings. However, IAHPR, as a whole or its Foundation, or its members do not take any responsibility for the completeness or correctness of data or references given by authors in publications and presentations at IAHPR events.

It is not within the remit of IAHPR or its Foundation, in particular, to seek clarification or detailed information from authors about data in submitted abstracts. Moreover, it is not within the scope of IAHPR and its committees to monitor compliance with any legal obligations, e.g., reporting requirements or regulatory actions.

Session 1 (9:15–10:45 CEST, Monday 05 September 2022):

To Pool or Not to Pool: Controlling for Multiple Subgroup-level Scales with LCA

Juan Marcos Gonzalez Sepulveda[1], F. Reed Johnson[1], Eric Finkelstein[2]

[1]Duke University, Durham, NC, USA; [2]Duke-NUS, Singapore, Singapore

Background: Poolability of preference data across subgroups is based on the compatibility of population-level means derived from subgroups. Adjusting for scale differences is required to fairly assess compatibility, but usual scale controls assume one scale factor within each subgroup of interest. This may not be sufficient if scale differences are strictly related to a group of respondents within subgroups.
Methods: We evaluate the poolability of data from a 5-country study looking at the relative importance of end-of-life care policies for caregivers. We compare a commonly-used approach to pool data with scale controls (1) with a novel approach that accounts for high-variance respondents based on task non-attendance, or the likelihood that respondents from each subgroup provided no meaningful information using latent-class analysis. We then evaluated poolability with results from both approaches.
Results: Results from the single-scale-controlled model showed that data from most countries should not be pooled. We found variation in the proportion of respondents who exhibited task non-attendance across countries, ranging between 19.1 and 45.1% ($P < 0.001$). Task non-attendance was significantly associated with data-validity failures. After controlling for variation in the distribution of high-variance respondents by country, we found that statistical tests support pooling data from all countries.
Conclusions: We find the proposed LCA provides a feasible way to help explain the proportion of respondents in a sample with large (small) variance (scale), and to use that information to control for their influence on the determination of poolability of data. Variation in the proportion of respondents who exhibit task non-attendance can lead to qualitatively different conclusions between the two methods used to determine poolability of choice data.

References:

[1] Hensher, David A., John M. Rose, and William H. Greene. "Combining RP and SP data: biases in using the nested logit 'trick'–contrasts with flexible mixed logit incorporating panel and scale effects." Journal of Transport Geography 16.2 (2008): 126–133.

Convergent Validity Between DCE and Other Stated Preference Methods: a Multi Study Comparison

Jorien Veldwijk[1], Esther de Bekker-Grob[1], Tommi Tervonen[2], Brett Hauber[3], Karin Groothuis-Oudshoorn[4]

[1]Erasmus School of Health Policy & Management, Rotterdam, the Netherlands; [2]Evidera, London, UK; [3]Pfizer, New York, USA; [4]University of Twente, Enschede, the Netherlands

Background: To start evaluating other methods beyond DCE for their applicability in assessing preferences for medical product lifecycle decision-making we aimed to assess the convergent validity of DCE compared to BWS case 1 and 2, swing weighting and Probabilistic Threshold Technique in four case studies: neuromuscular diseases (n = 140, DCE & BWS2), diabetes (n = 495, DCE & SW), myocardial infarction (n = 335, DCE & BWS 1), and rheumatoid arthritis (n = 982, DCE & PTT).
Methods: Results of the two methods were compared using a normalized preference measure for which confidence intervals (CIs) were estimated using non-parametric bootstrapping of 500 samples. Normalized preference measures comprised of mean relative attribute importance weights (NMD and diabetes studies), attribute uptake probability (MI study), or maximum acceptable risk (RA study).
Results: In all studies, attribute ranking showed similar patterns between the two methods for the most important attributes. In three out of four, the most important attribute was the same. However, significant differences were found in ranges of normalized preference measures across methods: 4.1–43.4 versus 8.9–24.7 for DCE and BWS2 in NMD; 3.8–49.7 versus 11.9–16.8 for DCE and SW in diabetes; 2.0–85.5 versus 0.2–69.0 for DCE and BWS case 1 in MI; − 3.5 to 49.2 versus 1.1–18.1 for DCE and PTT in RA.
Conclusions: Preferences differed significantly between methods implying limited convergent validity. The substantially larger ranges in normalized outcome measures in DCE compared to other methods, are likely due to differences in mechanics and bias related to the methods. Since none of the methods is considered the golden standard as true preferences are unknown, further studies are necessary to compare methods, determine internal validity and data quality and potentially measure external validity.

Improved Modelling of Interaction Effects in Discrete Choice Experiments

Marcel F. Jonker[1], Bas Donkers[2]

[1]Erasmus School of Health Policy & Management; [2]Erasmus School of Economics

Background: Discrete choice experiments (DCEs) are rarely analyzed with choice models that include a full set of two-way interactions between the attribute levels: the resulting model would be (too) difficult to interpret and the sampe size requirements (far) beyond what is feasible in applied research. Therefore, an alternative modelling approach is introduced that allows for interactions between the attributes—as opposed to interactions between the attribute's levels.
Methods: DCEs often comprise at least a subset of attributes for which monotonically increasing or decreasing preferences can be presumed, e.g. costs, benefits, risks, etc. Without imposing linear preferences, an 'optimal scaling' approach can be used to transform the levels of these attributes onto continuous latent scales, which can be interacted with each other and with the levels of categorical attributes. This results in a very parsimonious model specification.
Results: The proposed model with and without interactions is fitted on an existing dataset of N = 3699 respondents who each completed 16 EQ-5D-3L discrete choice tasks. As shown, the interactions between the attributes are straight-forward to interpret and their inclusion greatly improves statistical (WAIC) model fit statistics, while requiring 97% fewer parameters compared to a standard MIXL model with a full set of 2-way interactions between the included levels.
Conclusions: The proposed interaction model is parsimonious, produces estimates that are straight-forward to interpret, and accommodates the estimation of interactions in DCEs with more attractive and feasible sample size requirements. The model has one major disadvantage: it is not straight-forward to transform preferences for attributes with categorical levels onto a continuous latent scale.

Session 2 (11:00–12:30, CEST, Monday 05 September 2022):

Using Patient Treatment Beliefs to Inform DCE Designs

Charles Muiruri[1], Hayden Bosworth[1], Tianbei Zhang[1], Rita Kibicho[2], Rayanna Mboma[2], Shelby Reed[1], Reed Johnson[1], Juan Marcos Gonzalez[1]

[1]Duke University; [2]University of North Carolina, Chapel hill

Background: Intentional medication nonadherence is a conscious choice not to take medication that has been prescribed (1, 2). The association between medication nonadherence and adverse outcomes has been demonstrated in many observational studies (3, 4). However, patients' perceptions on the sensitivity of efficacy and side effects to nonadherence are poorly understood. Learning about these sensitivity thresholds can help inform relevant relative preferences that might make patients more vulnerable to nonadherence.
Methods: We developed a double-bounded contingent belief questionnaire for patients with hypertension and hiperlipidemia (5). Participants stated whether they thought efficacy and side effects associated with treatments for each condition were affected by an experimentally-controlled level of nonadherence. We used this information to estimate the relative sensitivity of outcomes to nonadherence. A DCE design was prepared to test if preferences overcome beliefs about outcomes associated with nonadherence.
Results: Results suggest that significant number of patients think efficacy is less sensitive to nonadherence than side effects. On average, patients expected they had to miss 36 times more doses to affect efficacy. These patients would be expected to manage side effects without sacrificing efficacy unless their preferences for these outcomes counteracted their beliefs. An 8-question design was generated to test whether patients consider efficacy to be at least 36 times more important than side effects.
Conclusions: It is possible to use conjoint analysis to understand patients' beliefs about how treatments "produce" health-related outcomes. Like early qualitative work, this information can be instrumental in constructing an efficient and principled DCE to evaluate health-related behaviors and preferences. In our case, this information will allow the evaluation of a connection between preferences and nonadherence for hypertension and hyperlipidemia medications.

References:
[1] Lehane E, McCarthy G. Intentional and unintentional medication non-adherence: A comprehensive framework for clinical research and practice? A discussion paper. International Journal of Nursing Studies. 2007/11/01/ 2007;44(8):1468–1477. https://doi.org/10.1016/j.ijnurstu.2006.07.010
[2] Cea-Calvo L, Marín-Jiménez I, de Toro J, et al. Different Associations of Intentional and Non-Intentional Non-Adherence Behaviors with Patient Experience with Healthcare and Patient Beliefs in Medications: A Survey of Patients with Chronic Conditions. Patient Prefer Adherence. 2020;14:2439–2450. https://doi.org/10.2147/PPA.S281985
[3] Walsh CA, Cahir C, Tecklenborg S, Byrne C, Culbertson MA, Bennett KE. The association between medication non-adherence and adverse health outcomes in ageing populations: A systematic review and meta-analysis. Br J Clin Pharmacol. 2019;85(11):2464–2478. https://doi.org/10.1111/bcp.14075
[4] Cutler RL, Fernandez-Llimos F, Frommer M, Benrimoj C, Garcia-Cardenas V. Economic impact of medication non-adherence by disease groups: a systematic review. BMJ Open. 2018;8(1):e016982. https://doi.org/10.1136/bmjopen-2017-016982
[5] Hanemann M, John Loomis, and Barbara Kanninen. Statistical efficiency of double-bounded dichotomous choice contingent valuation. American journal of agricultural economics 1991;73.4 1255–1263.

How to Decrease Social Desirability Bias in Stated Preference Data? Lessons Learned

Samare P.I. Huls[1], Job van Exel[1], Esther W. de Bekker-Grob[1]

[1]Erasmus School of Health Policy & Management (ESHPM), Erasmus University; Erasmus Choice Modelling Centre (ECMC), Erasmus University Rotterdam; Erasmus Centre for Health Economics Research (EsCHER), Erasmus University Rotterdam

Background: Discrete choice experiments (DCEs) have been used extensively to elicit preferences and inform healthcare decision-making. However, the hypothetical nature of choices induces socially desirable behaviour and endangers internal and external validity of DCEs. This study experimentally studied social desirability bias in DCEs and whether it can be mitigated using the cheap talk mitigation method.
Methods: Respondents (N = 1027) were randomly allocated to one of four study arms that differed in saliency of socially desirable behaviour and mitigation of this behaviour. The effect of social desirability bias and the cheap talk mitigation on internal validity was assessed by comparing respondent-reported characteristics, DCE results and the accuracy with which the models based on the stated preferences in the four arms predicted individual-level food choice in a holdout task.
Results: We found that social desirability bias, if present, was hardly affected by cheap talk mitigation. Respondent-reported characteristics, DCE results and prediction accuracy for the holdout task and real-world food choice did not strongly differ between study arms. Prediction accuracy for the holdout task was lowest in the default study arm (no manipulation). Prediction accuracy for real-world food choice was slightly better among respondents in the study arms exposed to cheap talk mitigation.
Conclusions: Considering the size of effects we conclude that social desirability bias was hard to detect and mitigate in this study. The effects we found indicate that cheap talk mitigation slightly improved external validity at a minimum cost of internal validity. Suggestions for future research include studying a context in which respondents are expected to show more socially desirable behaviour, using a different mitigation method, sampling from a real-world context and presenting a more salient opt-out.

The Impact of Violations of Expected Utility Theory on Choices Between Multiple Independent Risks

Juan Marcos Gonzalez Sepulveda[1], George Van Houtven[2], Shelby D. Reed[1], F. Reed Johnson[1]

[1]Duke University, Durham, NC, USA; [2]RTI International, Durham, NC, USA

Background: Use of preference information to infer risk tolerance for medical interventions has increased in recent years as a way to inform benefit-risk evaluations in regulatory and medical decision making. However, a framework for the quantification of preferences for multiple uncertain outcomes has not been formalized when choices do not comply with expected utility theory (EUT).

Methods: We develop a formal analytic framework for the measurement of preferences through choices under uncertainty with multiple independent risks. We also evaluate the implications of using various non-expected utility models on the framework. Finally, we apply the framework in a discrete choice experiment quantifying patient preferences for the benefits and risks of treatments for heart failure, including the risk of complications and death associated with treatment.
Results: Per the framework, we found that violations of EUT make preferences for uncertain outcomes conditional on other uncertain outcomes, not just nonlinear in probability. Nonlinearity in the risk of treatment complications influenced the estimated preferences for the chance of death ($P < 0.01$), but the chance of death did not affect the estimated preferences for the risk of complications ($P = 0.6$). These results indicate that commonly-used categorical models may misrepresent risk preferences.
Conclusions: We find supporting evidence that the framework correctly predicts effects found in data that captures choices under uncertainty through DCEs. Results from our framework imply that measures of risk tolerance derived from utility, such as maximum acceptable risk, must at least evaluate all relevant risks jointly if their effect on choices is expected to violate EUT.

Session 3 (15:15–16:45, CEST, Monday 05 September 2022):

Public Preferences in Organ Allocation: A Discrete Choice Experiment Regarding Distributive Justice

Carina Oedingen[1], Tim Bartling[1], Harald Schrem[2], Axel C Mühlbacher[3], Christian Krauth[1]

[1]Hanover Medical School, Germany, Center for Health Economics Research Hannover (CHERH); [2]Medical University Graz, Austria, Transplant Center Graz, Austria; [3]Hochschule Neubrandenburg, Germany, Duke University, USA

Background: There has been a persistent organ shortage, which forces priority setting in organ allocation to potential recipients. Because organ allocation is a highly normative decision task, it can be only legitimate if the general public is also involved in the decision making. Therefore, the aim was to assess public preferences for the allocation of deceased donor organs in Germany with the focus on ethical principles of distributive justice.
Methods: Based on systematic review and focus group discussions, six attributes each with two to four levels were selected: life years gained after transplantation, quality of life after transplantation, chance for a further donor organ offer, age, registered donor and individual role in causing organ failure. A fractional factorial design with a total of 104 choice sets (13 blocks with 8 choice sets) without opt-out was conducted. Data were analyzed using conditional logit, mixed logit and latent class.
Results: The final sample comprised 1028 respondents. Choice decisions were significantly influenced by all attributes except chance for a further donor organ offer. The conditional logit demonstrates that a good quality of life after transplantation, younger age and no individual role in causing organ failure had the greatest impact on choice decisions. Life years gained after transplantation and being a registered donor were less important.
Conclusions: The discrete choice experiment reveals that the probability of success in terms of a good quality of life after transplantation and a younger patients' age are most important in organ allocation for the public. Public preferences can help to inform policy to warrant socially responsible allocation systems and thus improve organ donation rates.

Is it Okay to Use Human Embryonic Stem Cells for Therapies? A Discrete Choice Experiment

Karin Schölin Bywall[1], Jennifer Drevin[1], Karin Groothuis-Oudshoorn[2], Jorien Veldwijk[3], Mats Hansson[1], Jennifer Viberg Johansson[1]

[1]Centre for Research Ethics and Bioethics, Uppsala University; [2]Health Technology & Services Research, University of Twente; [3]Julius Center for Health Sciences and Primary Care

Background: Human embryonic stem cells (hESC) based therapies may soon become a reality for Parkinson's disease (PD) (1, 2). The use of human embryos for therapies is associated with several ethical and legal issues (3, 4). This study assessed to what extent attitudinal questions and preferences resulting from a discrete choice experiment (DCE) can be used to inform the ethical and political debate.
Methods: 455 Swedish PD patients completed a DCE described by five attributes: type of treatment, aim of treatment, available knowledge of the different types of treatments, effect on symptoms, and risk for severe side effects. Latent class models were used to determine attribute-level estimates and preference heterogeneity (5). Relative importance and predicted uptake were calculated. Attitudinal questions were included related to moral status of embryos and handling left-over embryos.
Results: Three classes were identified. Class 1 reported a disutility for hESC, while class 2 and 3 preferred hESC over drug treatment, and focused on effect of symptoms and aim of the treatment. Respondents' preferences were associated with their experience with advanced treatment and side effects, but not their perceived moral status of the embryo. On average 65–69% of the respondents was predicted to accept hESC treatment depending on the level of associated severe side effects of treatments.
Conclusions: This study shows the added value of a DCE for investigating an ethically sensitive issue like using human embryonic stem cells for therapies. Preference heterogeneity was not impacted by the attitudinal questions. Current outcomes should be considered in the ethical and political debate on the use of hESC.
References
[1] Fan, Y., Winanto, and S.Y. Ng, Replacing what's lost: a new era of stem cell therapy for Parkinson's disease. Transl Neurodegener, 2020. 9: p. 2.
[2] Sugaya, K. and M. Vaidya, Stem Cell Therapies for Neurodegenerative Diseases. Adv Exp Med Biol, 2018. 1056: p. 61–84.
[3] Lo, B., & Parham, L. (2009). Ethical Issues in Stem Cell Research. Endocrine Reviews, 30(3), 204–213.
[4] Lo, B., & Parham, L. (2010). Resolving ethical issues in stem cell clinical trials: the example of Parkinson disease. Journal of Law Medicine & Ethics, 38(2), 257–266.
[5] Zhou, M., W.M. Thayer, and J.F.P. Bridges, Using Latent Class Analysis to Model Preference Heterogeneity in Health: A Systematic Review. Pharmacoeconomics, 2018. 36(2): p. 175–187.

An Evidence Base for Stated-Preference Research: Proving the Concept

Reed Johnson[1], Jui-Chen Yang[1], Meena Bewtra[2], Juan Marcos Gonzalez[1]

[1]Duke University; [2]University of Pennsylvania

Background: The maturation of health-preference research is indicated by the large number of published studies that have accumulated in some therapeutic areas. It is time to begin thinking of preference data in terms of evidence bases, similar to clinical data. We undertook

a proof-of-concept study to assess our ability to identify consensus risk-tolerance estimates from the body of evidence available on treatment preferences for Crohn's disease and ulcerative colitis.
Methods: We identified 22 published IBD preference studies, 7 of which reported discrete-choice-experiment (DCE) estimates useable for calculating maximum acceptable risk (MAR). Consensus published estimates were obtained by regressing MAR estimates on study-design characteristics. We also have obtained access to 5 original DCE datasets. The original data were pooled to estimate a serious-infection-scaled, MAR-space, data-fusion model. Assumptions were required to harmonize attribute definitions.
Results: Published results provided 314 individual MAR estimates. Including serious infection increased the estimate by 0.9% while malignancy decreased it by 0.8%. The pooled original datasets contained a total of 1366 respondents and over 21,000 choices. A treatment that improved IBD symptoms from moderate to remission, but with an annual cancer risk of 1% had an average annual serious-infection equivalent risk tolerance of 16.7%.
Conclusions: Stated-preference evidence bases in well-studied therapeutic areas can help establish consensus values for risk-tolerance measures, lend increased credibility for using stated-preference data to inform regulatory and clinical decision making, and enable leveraging previous research for benefit transfers to provide values in in the absence of sufficient time and funding for original studies, as well as help inform efficient, targeted new studies to fill identified gaps in the existing literature.

Session 4 (9:15–10:45, CEST, Tuesday 06 September 2022):

Classification of Functioning, Disability & Health: Validity of Preference-Weighted Scores

Christin Juhnke[1], Axel Christian Mühlbacher[2]

[1]Health Economics and Health Care Management, Hochschule Neubrandenburg, Neubrandenburg, Germany; [2]1_Health Economics and Health Care Management, Hochschule Neubrandenburg, Neubrandenburg, Germany/2_Duke Department of Population Health Sciences and Duke Global Health Institute, Duke University, 215 Morris St., Durham, North Carolina 27701, USA

Background: Indices, like the International Classification of Functioning Disability & Health (ICF) (ICF) are often used to measure outcomes for decision making (1). As other instruments it assigns equal weight to each item without distinguishing relevance (2). From a patients' perspective the validity of the results can be questioned. The objective is to examine the extent to which a preference-weighted score of core elements of the ICF differs from unweighted scores currently used in clinical decisions.
Methods: Three best–worst scaling experiments are used to value ICF dimensions: movement, neglect, activities (3). Stroke patients and members of the public are recruited. The ICF "tariff" is generated by converting ICFs percentual generic qualifiers of impairments to a unidimensional index. The preference weights for levels within each dimension are produced on a 0–1 scale, in which the most desirable level is assigned 1.
Results: N = 306 participants were recruited in May/June 2022. There is evidence of divergent validity of preference-weighted and unweighted scores for the ICF based on the intraclass correlation coefficient. Unweighted scores in the dimensions body function, activity and perception differ considerably from the preference-weighted scores derived from BWS experiments in general (e.g., ICC_movement: 0.889, ICC_neglect: 0.482) as well as for various hypothetical health states constructed based on ICF states.
Conclusions: Body functions have effects on activities, and these effect on health-related quality of life. This raises the question how the value is measured and considered in decision-making? The results reveal that ICF body functions or activities are not equally weighted by those affected.
This fosters discussions on the differences of preference-weighted scores and simple additive models to develop a patient-centered classification of impairments and therapy goals.

References

(1) World Health Organization (2001). International classification of functioning, disability and health: ICF. Geneva, World Health Organization.
(2) Appleby J, Devlin N, Parkin D. Using Patient Reported Outcomes to Improve Health Care. Wiley, 2016
(3) Mühlbacher, A.C., Kaczynski, A., Zweifel, P., Johnson, F.R., Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. Health economics review, 2015. 6(1): p. 1–14.

Online Elicitation of Personal Utility Functions (OPUF): An Open, Modular Health Valuation Platform

Paul Schneider[1], John Brazier[1], Ben van Hout[2], Nancy Devlin[3],

[1]University of Sheffield, Sheffield, UK; [2]University of Sheffield, Sheffield, UK; Open Health, York, UK; [3]University of Melbourne, Melbourne, Australia

Background: Commonly used preference elicitation methods, such as TTO or DCE, usually require data from hundreds of participants. Conducting health valuation studies thus becomes time and resource intensive, and eliciting preferences from patient, alongside HTA, for example, is often deemed infeasible altogether. This severely limits the availability of relevant (patient) preference information to decision makers. Here, we report on the development of OPUF; a new approach for valuing health and well-being.
Methods: OPUF allows constructing value functions for small groups and even on the individual person level. The approach combines different compositional preference elicitation techniques into a new type of online survey. It broadly consists of three steps: dimension weighting, level rating, and anchoring. Demo OPUF surveys are available at: https://valorem.health.
Results: We successfully piloted the OPUF approach for the EQ-5D-5L in samples of the general population in the UK (n = 1000) and Germany (n = 500), and in patients with rheumatoid arthritis (n = 90). The median completion time was 8–14 min. After excluding participants who skipped one or more valuation steps, we were able to construct a preference function for each participants.These personal utility functions predicted participants' choices in hold-out DCE tasks with an accuracy of about 80%.
Conclusions: Although OPUF is still under development, early results are promising, and we see several potential future applications. Most notably, OPUF could be used to elicit preferences from small groups of patients (e.g. patients with rare diseases), when other established methods seem infeasible. We now seek to make the approach available to others, and started developing a modular, open source online platform, which will allow researchers to design, launch, and analyse online health valuation studies.

References:

https://valorem.health
https://wellcomeopenresearch.org/articles/7-14/v1
https://link.springer.com/article/10.1007/s10198-018-0993-z
https://bitowaqr.github.io/files/opuf_uk.pdf

A Novel Approach to Computing Preference Estimates for Different Treatment Pathways in Oncology

Kathleen Beusterien[1], Oliver Will[1], Susan McCutcheon[2], Emuella Flood[3], deMauri Mackie[1], Stella Mokiou[2]

[1]Cerner Enviza, Malvern, PA; [2]AstraZeneca, Cambridge, UK; [3]AstraZeneca, Gaithersburg, MD

Background: Treatment pathways in cancer are frequently complex, with differing treatment sequencing and therapies offering varying benefit to risk. Moreover, pathways may be fixed or flexible in allowing for escalation or de-escalation of treatment depending upon interim outcomes. We sought to develop a methodology capable of estimating preferences for an overall treatment pathway with sequential treatments, using early breast cancer (eBC) patient preference data from Germany, Italy, and Japan.
Methods: Patients completed an online discrete choice experiment to assess preferences for 8 key eBC treatment pathway attributes (1). Hierarchical Bayesian modelling was used to calculate preference weights (PW). PW for hypothetical pathways were estimated by summing the respective PW for efficacy, flexible or fixed, and duration; and for administration route/schedule and adverse event (AE) risks, which were time adjusted by multiplying each weight by the proportion of time spent on a selected treatment.
Results: 452 patients took part (2). The reference case (flexible pathway; oral therapy for 18 months [mos]; 86% 3-year event-free survival; grade $\geq$ 3 AE risk of 24%) had a mean PW (mPW; [95% CI] of 1.63 [1.50, 1.76]. Decreasing therapy from 18 mos to 12 mos increased the mPW to 2.07 [1.94, 2.20]. Switching to a two-phase pathway with therapy for 6 mos followed by a different therapy for 12 mos, each with different risks of grade $\geq$ 3 AE risk of (77% and 24%, respectively) decreased the mPW to 0.71 [0.58, 0.83].
Conclusions: Using eBC as a reference case, we show how this novel methodology expands beyond traditional preference elicitation by accounting for patient preference throughout a sequence of treatments which may vary in their durations, risks, and mode of administration. A limitation is that the preference weights are based on a cross-sectional point in time from patients with varying treatment experiences, which can be further explored in sensitivity and subgroup analyses.

References:
[1] Gennari A et al. Qual Life Res 2021;30:1–117 (abstract 1006)
[2] Gennari A et al. Ann Oncol 2022;33:S154–5 (abstract 70P)

Session 5 (11:00–12:30, CEST, Tuesday 06 September 2022):

Participatory Value Evaluation (PVE): A New Preference-Elicitation Method for Health Decision-Making

Sander Boxebeld[1], Job van Exel[1], Niek Mouter[2]

[1]Erasmus University Rotterdam, Rotterdam, the Netherlands; [2]Delft University of Technology, Delft, the Netherlands

Background: As an alternative to established preference elicitation methods like discrete choice experiments (DCE), Participatory Value Evaluation (PVE) has been introduced in other fields (1–3) and was recently first applied in the field of health. In a PVE, participants select a portfolio of policy options described by attributes, within a constraint. This paper aims to illustrate PVE's potential in healthcare decision-making and to position it relative to established preference elicitation methods.
Methods: We describe the PVE-method and its theoretical background and illustrate its potential by discussing the first three published studies applying PVE in the health domain (4–6) Subsequently, we position PVE relative to several other preference elicitation methods by comparing the structure and focus of choice tasks.
Results: The portfolio-based choice task in a PVE allows participants to consider a set of actual policy alternatives in relation to each other, but may impose higher cognitive burden on participants and is less efficient than methods including multiple choice tasks. The constraint in the PVE choice task forces participants to make trade-offs within the restrictions faced by policymakers. A flexible budget constraint additionally allows participants to trade-off their private income with public spending.
Conclusions: PVE seems a promising method for involving preferences of citizens or patients in healthcare decision-making. PVE is especially distinct from other methods in its ability to simultaneously elicit public preferences for policy alternatives and for the trade-off between public and private spending. This may be particularly useful in the context of healthcare decision-making. Further research is required into the feasibility for different groups of citizens and validity of PVE.

References:
[1] Mouter, N., Koster, P. and Dekker, T. (2022). Participatory value evaluation for the evaluation of flood protection schemes. Water Resources and Economics, 36, 100188
[2] Mouter, N., Koster, P. and Dekker, T. (2021). Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. Transportation Research Part A: Policy and Practice, 144, 54 – 73
[3] Mouter, N., Shortall, R.M., Spruit, S.L. and Itten, A.V. (2021). Including young people, cutting time and producing useful outcomes: Participatory value
[4] Mulderij, L.S., Hernández, J.I., Mouter, N., Verkooijen, K.T. and Wagemakers, A. (2021). Citizen preferences regarding the public funding of projects promoting a healthy body weight among people with a low income. Social Science & Medicine, 280, 114015
[5] Mouter, N., Hernández, J.I. and Itten, A.V. (2021). Public participation in crisis policymaking. How 30,000 Dutch citizens advised their government on relaxing COVID-19 lockdown measures. PLoS ONE, 16(5), e0250614
[6] Rotteveel, A.H., Lambooij, M.S., Over, E.A.B. et al. (2022). If you were a policymaker, which treatment would you disinvest? A participatory value evaluation on public preferences for active disinvestment of health care interventions in the Netherlands. Health Economics, Policy and Law [article in press]

Multi-Dimensional Thresholding for Eliciting Multi-Attribute Treatment Preferences

Tommi Tervonen[1], Sebastian Heidenreich[1], Douwe Postmus[2]

[1]Evidera, London, UK; [2]Department of Epidemiology, University Medical Center Groningen, University of Groningen, The Netherlands

Background: Thresholding technique is a viable method for health preference studies that need individual level preferences or have limited sample sizes. However, application of thresholding to multi-dimensional context with more than a single trade-off is not trivial and can be implemented using a range of approaches that differ in

complexity. We aimed to evaluate three variants of the multi-dimensional thresholding method in simulations.
Methods: Each variant included an initial ranking of the attribute scale swings followed by thresholding on individual trade-offs, in rank order, using bi-sectioning based on (i) level range, (ii) unconditional utility, and (iii) conditional utility. We conducted simulations for hypothetical problems with 3–5 attributes (i.e. 2–4 trade-offs) and 2–5 thresholding questions per trade-off post-ranking. For each parameter combination, 1000 simulations were run with random target trade-off (weight) vectors.
Results: Lowest median Euclidean distance to the target weight vector was 0.004 for conditional utility with 5 attributes and 5 elicitation questions per trade-off. Inter-quartile ranges varied between 0.004 and 0.007 (level range, 5 attributes, 5 questions) and 0.028–0.064 (level range, 3 attributes, 2 questions). More outliers (Euclidean distance > 0.1) were observed for conditional utility (2.52% of simulations) than for level range (0.95%) or conditional utility (0.88%) variants.
Conclusions: Multi-dimensional thresholding achieves sufficient precision with low number of elicitation questions. Choice of algorithm for constructing elicitation questions has limited impact on precision of the results and all variants are expected to perform adequately in practical applications with up to 5 attributes.

Modeling Health Preferences Using Machine Learning: Preliminary Evidence from a DCE on HIV Testing

Tengjie Tang[1], Chen Liang[2], Ethan Fang[1], Nathan Thielman[1], Jan Ostermann[2]

[1]Duke University; [2]University of South Carolina

Background: Standard methods for the analysis of DCE data are subject to various constraints and do not easily support the inclusion of large numbers of respondent characteristics as covariates. We sought to evaluate the feasibility and predictive performance of machine learning (ML) methods for modeling DCE choices among HIV testing alternatives as a function of respondent characteristics and characteristics of the alternatives shown in each choice task.
Methods: We implemented ML algorithms, including Random Forest (RF) and XGBoost on existing DCE data (n = 740, 12 choices from 3 alternatives each) to classify whether respondents preferred or did not prefer an alternative. In three specifications, ML predicted the (1) best-of-three, (2) best vs. 2nd-best, or (3) best vs. 3rd-best alternatives. ML models were tuned using 5-fold cross validation (CV). Matthews correlation coefficient (MCC) and SHAP were used as evaluation metrics.
Results: RF had the best performance, outperforming other ML and mixed logit models. RF predicted the best of 3 alternatives with in-sample MCC = 0.51 (CV MCC = 0.35, mixed logit in-sample MCC = 0.36); RF predicted the best vs. 2nd-best alternative with MCC = 0.53 (CV MCC = 0.3, mixed logit in-sample MCC = 0.28); RF predicted the best vs. 3rd-best alternative with MCC = 0.68 (CV MCC = 0.53, mixed logit in-sample MCC = 0.50). Respondent characteristics had lower SHAP feature importance than alternative attributes.
Conclusions: We demonstrated the feasibility and acceptable performance of tested ML methods for modeling DCE choices and predicting HIV testing preferences. Attributes of the evaluated alternative and the other alternatives shown in the same choice task were more important than respondent characteristics. The effect of (1) including additional variables and (2) of within vs. out-of-sample (CV) predictions on the relative performance of ML methods vs. mixed logit models should be explored.

References:

[1] J. Ostermann, B. P. Flaherty, D. S. Brown, B. Njau, A. M. Hobbie, T. B. Mtuy, M. Masnick, A. C. Mühlbacher, and N. M. Thielman, "What factors influence hiv testing? modeling preference heterogeneity using latent classes and class-independent random effects," Journal of Choice Modelling, vol. 40, p. 100305, 2021.
[2] J. Ostermann, B. Njau, A. Hobbie, T. Mtuy, M. Maxnick, D. Brown, A. Mühlbacher, and N. Thielman, "Divergent preferences for enhanced hiv testing options among high-risk populations in northern tanzania: A short report," AIDS care, 2022 (forthcoming).
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.
[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
[5] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 2522–5839, 2020.

Spotlight Session 6 (13:30–15:30, CEST, Tuesday 06 September 2022):

The Garbage Class Mixed Logit Model: Accounting for Low-Quality Response Patterns in DCEs

Marcel Jonker[1]

[1]Erasmus School of Health Policy & Management; Erasmus Choice Modelling Centre

Background: The aim of this presentation is to introduce the garbage class mixed logit (MIXL) model as a convenient alternative to manually screening and accounting for respondents with low data quality in discrete choice experiments (DCEs), e.g. based on internal validity or statistical validity tests like the root likelihood (RLH) test.
Methods: Garbage classes are typically used in latent class logit analyses to designate or identify group(s) of respondents with low data quality. Yet the same concept can be applied to MIXL models as well. In its most basic form, the model has 2 latent classes: the first represents the standard MIXL model that one would normally fit (e.g. when computing RLH statistics), whereas the second class represents a so-called 'garbage class' in which respondents are assumed to make arbitrary choices.
Results: Based on a re-analysis of 4 different DCEs, MIXL models with a garbage class can achieve the same (or better) accuracy as manually screening for respondents with low data quality. However, the garbage class model has the advantage of producing MIXL estimates that are purged from the influence of respondents with low-quality response patterns and providing estimates of the number of respondents with low-quality response patterns in the dataset, without the need for statistical cut-off values.
Conclusions: Although less versatile than the combination of standard MIXL estimates with sensitivity analyses, the proposed garbage class MIXL model automatically accounts for respondents with low-quality response patterns and provides an objective quantification of DCE data quality that is consistent with the underlying theoretical framework of DCEs. Moreover, it is straight-forward to extend the model with two or more MIXL classes.

Wrong or Unexpected? Using Internal Validity Tests to Understand Stated Preference Survey Results

Deborah A Marshall[1], Karen V MacDonald[1], F Reed Johnson[2]

[1]University of Calgary, Calgary, Alberta, Canada; [2]Duke University, Durham, NC, USA

Background: Internal validity tests (IVTs) are used in discrete-choice experiments (DCEs) to check choice logic, response consistency, and extent respondents accept trade-offs (1). However, there are no criteria for how many IVT failures would classify a respondent as having unacceptably poor quality data. Further, respondents may have reasonable explanations for their choices (2). We conducted an IVT experiment in a DCE to understand why respondents fail IVTs and impact of failures on preference estimates.
Methods: We conducted a DCE with 4 attributes, 12 experimental choice tasks, and 2 constructed IVT choice tasks (stability test and within-set dominated pairs test). Respondents who failed these IVT were asked to explain their choices. We used a previously developed IVT tool (1) to conduct additional IVTs and analyze dominance patterns. Latent-class analysis (LCA) was used to identify taste heterogeneity. We also used forced known-class assignments to evaluate possible dominance-acceptability thresholds.
Results: Of 201 respondents, 18% failed $\geq$ 1 constructed IVT. About half of those who failed provided detailed explanations for their choices. Using a threshold of 75% dominance on ordered attributes, 62% failed the dominance test. In a 4-class LCA, we constrained those who passed this test to class 1 and those who failed to class 4; all respondents were allowed in classes 2 and 3 (unconstrained classes). For those who failed, membership probability for class 2 or 3 was 0.86, and 0.73 for those who passed.
Conclusions: Failing IVTs could be explained by reasons other than low engagement. Dominant responses could be a result of low engagement or strong preference for one attribute. We found most respondents who failed the dominance test still provided sufficient trade-offs to classify them evenly between two unconstrained latent classes. IVT failures should be interpreted as unexpected responses warranting further inquiry. Including follow-up questions in surveys could yield insights about stated preferences.

References:
[1] Johnson FR, Yang JC, Reed SD. The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments. Value Health. 2019; 22(2):157–160.
[2] Jonker MF, Roudijk B, Maas M. The Sensitivity and Specificity of Repeated and Dominant Choice Tasks in Discrete Choice Experiments. Value Health. 2022.

Within-Set Dominated Pair for Validity Testing: Unexpected Results from a Discrete Choice Experiment

Melanie Brinkman[1], Christian Krauth[1], Bernt-Peter Robra[2], Maren Dreier[1]

[1]Institute for Epidemiology, Social Medicine and Health Systems Research, Hannover Medical School, Hannover, Germany; [2]Institute of Social Medicine and Health Systems Research, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Background: Good research practices for discrete choice experiments (DCE) in health include testing the internal validity of the choice data (1,2). Therefore, in our DCE on preferences of colorectal cancer screening tests we included a within-set dominated pair where one alternative was clearly superior to the other. However, an unexpectedly high proportion of respondents (476/1142; 'irrational' (IR)) chose the dominated alternative, which needs further methodological consideration and discussion.
Methods: DCE with generic two-alternative choice sets and 6 attributes (mortality, incidence, complications, preparation, need for transportation, follow-up), each with 3 levels. Participants completed 8 choice tasks and in addition the within-set dominated pair for validity testing. A stratified random sample (n = 5000) of 50, 55, and 60 year olds was invited to participate (June 2020). Preferences were analyzed using conditional logit stratified by responses to the internal validity test (3, 4).
Results: Of 1282 questionnaires received, 1142 were included. 'Rational' (R) (n = 666) and IR (n = 476) respondents differed in sociodemographics, screening history, screening intention, health literacy, and certainty in choices made. R respondents valued cancer-specific mortality and incidence most, IR respondents preparation and accompaniment home. Contrary to a priori expectations, higher effort levels were preferred for bowel cleansing (R and IR) and accompaniment home (IR).
Conclusions: IR respondents showed different preferences than those who answered the internal validity test as expected. Reasons for this 'irrational' appearing behavior remain unclear. It can reflect either unknown but 'rational' considerations or shortcomings in the design. To clarify whether or not 'irrational' appearing responses are preference-based, qualitative research or open-ended questions about motivations behind choice behavior should be considered in future DCE surveys (5).

References:
[1] Bridges JFP, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health-a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health. 2011; 14:403–13. https://doi.org/10.1016/j.jval.2010.11.013.
{2] Johnson FR, Yang J-C, Reed SD. The internal validity of discrete choice experiment data: a testing tool for quantitative assessments. Value Health. 2019; 22:157–60.
[3] Hauber AB, González JM, Groothuis-Oudshoorn CGM, Prior T, Marshall DA, Cunningham C, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value Health. 2016; 19:300–15. https://doi.org/10.1016/j.jval.2016.04.004.
[4] Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences. Health Econ. 2006; 15:797–811. https://doi.org/10.1002/hec.1104.
[5] Ryan M, Watson V, Entwistle V. Rationalising the 'irrational': a think aloud study of discrete choice experiment responses. Health Econ. 2009; 18:321–36. https://doi.org/10.1002/hec.1369.

Evaluating External Validity of a Discrete Choice Experiment: Preferences for Labor Pain Medicine

Semra Ozdemir[1], Eric Finkelstein[1], Prateek Bansal[2], Juan Marcos Gonzalez[1]

[1]Duke-NUS Medical School; [2]National University of Singapore

Background: The external validity of the Discrete Choice Experiments (DCEs) has been questioned because of the hypothetical nature of the method, and has been understudied in health care research due to the limited opportunities to test it (1). In this study, we aim to test the external validity of the DCE method by comparing predicted choices from a DCE survey with actual choices in real life.

Methods: The case study is a DCE conducted to assess preferences for epidural analgesia among 248 women who were admitted to a maternal institution for childbirth. We used both mixed logit (MXL) and latent class logit (LCM) models allowing for preference heterogeneity. We compared the predicted probabilities of epidural with the choices made in real life at the sample and individual levels. We also calculated the positive predictive value (PPV) and the negative predictive value (NPV).

Results: 80% of the subjects used epidural in real life while we predicted an epidural uptake of 64% from MXL and 59% from LCM at the sample level. The proportions of women who had concordance between their predicted choices and real-life choices were 56% and 50% from MXL and LCM, respectively. The PPV and NPV were 87% and 33%, respectively for MXL while they were 86% and 28%, respectively for LCM. None of the personal characteristics explained the concordance between the predicted and real-life choices.

Conclusions: Our models were better at predicting at the sample level than at the individual level, and at predicting those who chose epidural (PPV) than predicting those who did not choose epidural (NPV) in real life. Our study did worse at predicting than those from previous studies with a concordance ranging from 64 to 83% (2–5). Hot-cold empathy gap may explain why our DCE under-predicted the choice of epidural compared to real-life choices.

References:

[1] Quaife, M., et al., How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity. The European journal of health economics, 2018. 19(8): p. 1053–1066.

[2] Mohammadi, T., et al., Testing the external validity of a discrete choice experiment method: an application to latent tuberculosis infection treatment. Value in Health, 2017. 20(7): p. 969–975.

[3] Salampessy, B.H., et al., The predictive value of discrete choice experiments in public health: an exploratory application. The Patient-Patient-Centered Outcomes Research, 2015. 8(6): p. 521–529.

[4] Lambooij, M.S., et al., Consistency between stated and revealed preferences: a discrete choice experiment and a behavioural experiment on vaccination behaviour compared. BMC medical research methodology, 2015. 15(1): p. 1–8.

[5] Linley, W.G. and D.A. Hughes, Decision-makers' preferences for approving new medicines in wales: a discrete-choice experiment with assessment of external validity. Pharmacoeconomics, 2013. 31(4): p. 345–355.