



From Individual to Population Preferences: Comparison of Discrete Choice and Dirichlet Models for Treatment Benefit-Risk Tradeoffs

Tommi Tervonen , Francesco Pignatti, and Douwe Postmus

Introduction. The Dirichlet distribution has been proposed for representing preference heterogeneity, but there is limited evidence on its suitability for modeling population preferences on treatment benefits and risks. **Methods.** We conducted a simulation study to compare how the Dirichlet and standard discrete choice models (multinomial logit [MNL] and mixed logit [MXL]) differ in their convergence to stable estimates of population benefit-risk preferences. The source data consisted of individual-level tradeoffs from an existing 3-attribute patient preference study ($N = 560$). The Dirichlet population model was fit directly to the attribute weights in the source data. The MNL and MXL population models were fit to the outcomes of a simulated discrete choice experiment in the same sample of 560 patients. Convergence to the parameter values of the Dirichlet and MNL population models was assessed with sample sizes ranging from 20 to 500 (100 simulations per sample size). Model variability was also assessed with coefficient P values. **Results.** Population preference estimates of all models were very close to the sample mean, and the MNL and MXL models had good fit (McFadden's adjusted $R^2 = 0.12$ and 0.13). The Dirichlet model converged reliably to within 0.05 distance of the population preference estimates with a sample size of 100, where the MNL model required a sample size of 240 for this. The MNL model produced consistently significant coefficient estimates with sample sizes of 100 and higher. **Conclusion.** The Dirichlet model is likely to have smaller sample size requirements than standard discrete choice models in modeling population preferences for treatment benefit-risk tradeoffs and is a useful addition to health preference analyst's toolbox.

Keywords

decision analysis, health preference elicitation, patient choice modeling, pharmacoepidemiology

Date received: October 31, 2018; accepted: August 5, 2019

Preference studies are increasingly being used to support health policy decision making with regulatory agencies recently expressing interest in preference-based benefit-risk assessment.^{1–3} Discrete choice experiments (DCEs) are the most commonly used method for eliciting benefit-risk tradeoffs in the health domain.⁴ In a DCE, benefit-risk tradeoffs are inferred from a series of choice questions in which participants are asked to choose between 2 or more hypothetical treatment profiles. The utility that a participant obtains from a treatment profile is assumed to be a random variable whose expected value is expressed as a function of the attribute levels that constitute that treatment profile. The regression

coefficients of this function are the parameters of interest for the DCE and can be used to calculate attribute

Evidera, London, UK (TT); European Medicines Agency, Amsterdam, the Netherlands (FP); Department of Epidemiology, University of Groningen, University Medical Center Groningen, the Netherlands (DP). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article.

Corresponding Author

Tommi Tervonen, Evidera, The Ark, 2nd Floor, 201 Talgarth Road, London W6 8BJ, UK (Tommi.Tervonen@evidera.com).

weights that express the marginal rate of substitution between 2 attributes.

Because of the limited amount of information that is obtained with each discrete choice question, DCEs often require hundreds of answers for estimating the preference parameters (i.e., the benefit-risk tradeoffs, possibly conditional on a set of explanatory covariates) sufficiently accurately.⁵ Moreover, the maximum number of attributes respondents can handle in DCEs is limited; the exact number is context dependent,⁶ but most recently published health DCEs have used between 4 and 9 attributes.⁷ Finally, although more complex statistical models allow distinguishing and characterizing preference heterogeneity,⁸ DCEs rarely allow estimating individual-level utility functions with high precision.⁹

Other preference elicitation and modeling methods have been developed to overcome these challenges. Instead of assuming that utility is a latent variable, multicriteria decision analysis is based on normative models of rational choice that state that a subject's preference structure can be represented by means of a utility function if that subject's choice behavior satisfies certain basic rationality axioms, such as completeness and transitivity. Several direct valuation methods have been developed to elicit the parameters of this function, which has an additive structure when the attributes under consideration are preferentially independent for the decision maker. The parameters of interest for the additive value model are the attribute weights and the marginal gain or loss in utility from increasing attribute values (i.e., the so-called partial value or utility functions).

Swing weighting and other direct valuation methods can handle a larger number of attributes, and their questioning procedures are designed in such a way that they completely identify an individual's utility function. However, when direct valuation methods are used, and the analyst wants to generalize from the sample to the population, a statistical model for the data-generating process needs to be specified. Various authors have proposed using the Dirichlet distribution for modeling the distribution of the attribute weights in the population.¹⁰⁻¹² The Dirichlet distribution is particularly compelling for this purpose, given its support is the simplex (i.e., the full feasible space of attribute weights when they are normalized to sum to unity). However, there is limited empirical evidence on the use of the Dirichlet distribution to model population preferences, including an understanding of the convergence of the parameter estimates with sample sizes commonly encountered in health preference studies.

This article aims to fill this evidence gap by evaluating the use of the Dirichlet distribution for modeling

population preferences. We compare the Dirichlet distribution to the multinomial logit model (MNL) commonly used for modeling preferences as captured with a DCE, by conducting computational experiments with data collected in a previous preference study. We also fit a mixed logit (MXL) model to the data and discuss the differences between the MXL and Dirichlet approaches.

Methods

We conducted a simulation study to compare how the Dirichlet and MNL models differ in their convergence to stable estimates of the population preferences.

We based our computational experiments on an existing study¹³ that used an online questionnaire with choice-based matching questions to elicit the preferences of 560 patients with multiple myeloma for hypothetical cancer treatments. The treatments were described in terms of the following 3 attributes: probability of being progression free for 1 year or longer (index: 1; levels: 50%, 60%, 70%, 80%, and 90%), risk of moderate but chronic toxicity (index: 2; levels: 45%, 55%, 65%, 75%, and 85%), and risk of severe toxicity (index: 3; levels: 20%, 35%, 50%, 65%, and 80%). The source data consisted of a set of $N = 560$ weight vectors (1 for each patient) that were derived from the patients' responses to the choice-based matching questions. Using these real data, instead of simulated data, may provide better evidence on the methods' convergence in a realistic setting.

For the Dirichlet model, the utility that a random patient i obtains from a hypothetical treatment j with attribute values $x_j = (x_{j1}, x_{j2}, x_{j3})$ was specified as

$$U_{ij} = w_{i1} \left(\frac{x_{j1}}{90 - 50} \right) - w_{i2} \left(\frac{x_{j2}}{85 - 45} \right) - w_{i3} \left(\frac{x_{j3}}{80 - 20} \right).$$

Here, the attribute weights $w_i = (w_{i1}, w_{i2}, w_{i3})$, which are nonnegative and normalized to sum to unity, are Dirichlet distributed with density $f_D(w)$. To estimate $f_D(w)$, we fit a Dirichlet regression model directly to the attribute weights in the source data. We refer to this fitted distribution as $\hat{f}_D(w)$.

For comparison purposes, we also fit an MNL model and MXL model to the outcomes of a simulated DCE in the same sample of 560 patients. To obtain discrete choice data sets for fitting the 2 models, we simulated a DCE with the following design. First, we generated an orthogonal design using the L^{ma} method¹⁴ with 2 choice alternatives and 5 levels for each of the 3 attributes. Then, we filtered out questions in which 1 of the choice alternatives was dominated (i.e., would have higher risk

of both toxicities and lower probability of 1-year progression-free survival). This resulted in a set of 16 possible discrete choice questions. Next, we simulated the individual patient responses based on the behavioral assumption that, for any given question, patients choose the alternative that provides the greatest utility. The utility that patient i obtained from alternative j in choice question k was generated according to the following equation:

$$\tilde{U}_{ijk} = w_{i1} \left(\frac{x_{jk1}}{90 - 50} \right) - w_{i2} \left(\frac{x_{jk2}}{85 - 45} \right) - w_{i3} \left(\frac{x_{jk3}}{80 - 20} \right) + \epsilon_{ijk}.$$

Here, ϵ_{ijk} is a Gumbel distributed error term that was added to the utility values of the choice alternatives in each question to make the resulting choice behavior consistent with the assumptions underlying the MNL and MXL models.^{15–17} The attribute weights in this equation were directly taken from the source data.

To determine a suitable scale value β for the Gumbel distributed error term, we conducted for each value of $\beta \in \{0.1, 0.2, \dots, 1.0\}$ a set of 1000 experiments in which each of the $N = 560$ patients were simulated to answer all 16 discrete choice questions. These results indicated that the expected Euclidean distance between the sample mean of the attribute weights (i.e., the arithmetic mean of the attribute weights of all 560 patients in the sample) and the normalized attribute weights calculated from the regression coefficients of the fitted MNL models was minimal when the Gumbel scale value was set equal to $\beta = 0.3$ (see the results in the Supplementary Material). This scale value was therefore used for further simulations of the MNL and MXL models.

To fit the MNL and MXL models to the outcomes of the simulated DCE in the sample of 560 patients, we used linear models with continuous level encoding to estimate only the coefficients that express marginal rates of substitution between the attributes:

$$\tilde{U}_{ijk} = \tilde{w}_{i1}x_{jk1} + \tilde{w}_{i2}x_{jk2} + \tilde{w}_{i3}x_{jk3} + \tilde{\epsilon}_{ijk}.$$

Here, the vector of preference weights \tilde{w}_i is either fixed (MNL) or independent, normal-distributed (MXL) and $\tilde{\epsilon}_{ijk}$ Gumbel distributed with $\beta = 1$. We refer to these fitted distributions (deterministic distribution in case of MNL) as $\hat{f}_{MNL}(\tilde{w})$ and $\hat{f}_{MXL}(\tilde{w})$, respectively. The normalized attribute weights w_i can be obtained from the preference weights \tilde{w}_i by multiplying the latter with the attribute scale variation and then applying rescaling so that they sum to unity. To achieve comparability of the

Table 1 Estimates of Population Means of Normalized Attribute Weights Based on the Sample Mean and the Fitted Dirichlet, MNL, and MXL Models, as Well as the Maximum Acceptable Risks of AEs to Increase the Probability of 1-Year PFS by 1%

Attribute	Dirichlet		MNL		MXL	
	Sample Mean (95% CI)	Normalized Weight (95% CI)	Mean (SE)	Normalized Weight (95% CI)	Mean (SE)	Normalized Weight (95% CI)
1-year PFS	0.54 (0.52–0.55)	0.52 (0.50–0.54)	0.033 (0.001)	0.54 (0.52–0.56)	0.047 (0.002)	0.54 (0.52–0.56)
Moderate AEs	0.14 (0.13–0.15)	0.17 (0.16–0.18)	–0.009 (0.001)	0.13 (0.11–0.16)	–0.012 (0.001)	0.13 (0.11–0.16)
Severe AEs	0.32 (0.30–0.34)	0.31 (0.30–0.33)	–0.015 (0.001)	0.33 (0.31–0.35)	–0.019 (0.001)	0.33 (0.31–0.35)
MAR (SE) moderate AEs	3.79% (0.16)	3.06% (0.11)	4.15% (0.46)		4.00% (0.42)	
MAR (SE) severe AEs	2.49% (0.11)	2.48% (0.11)	2.44% (0.12)		2.47% (0.12)	
Adjusted McFadden's R^2			0.12		0.13	

AE, adverse event; CI, confidence interval; MNL, multinomial logit; MXL, mixed logit; PFS, progression-free survival; SD, standard deviation; SE, standard error; MAR, maximum acceptable risk.

^aDirichlet distribution SEs are on log scale. The 95% confidence intervals are [2.711, 3.239] (PFS), [0.890, 1.055] (moderate AEs), [1.642, 1.956] (severe AEs).

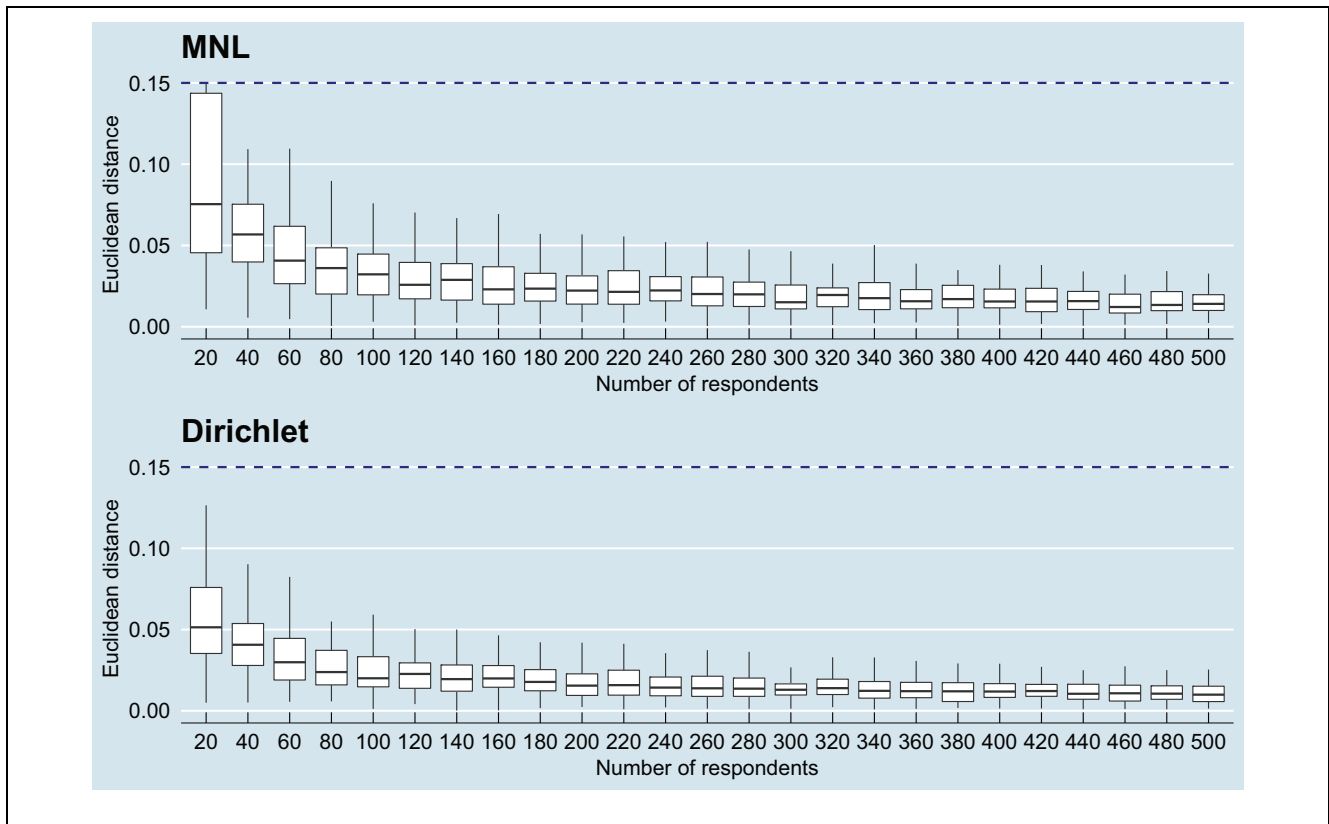


Figure 1 Box plots of convergence of the multinomial logit (MNL; top) and Dirichlet (bottom) models to the fitted population models with varying sample sizes; the dashed blue line indicates the Euclidean distance 0.15 that has been used to truncate the data set.

results of the different models, all comparisons were made at the level of the normalized attribute weights.

To assess the goodness of fit of the 3 models, the mean normalized attribute weights from the fitted models were compared with the sample mean of the attribute weights in the source data. Standard errors and 95% confidence intervals (CIs) for the mean attribute weights of the MNL and MXL models were obtained using the delta method.¹⁸ The 95% CI for the sample mean and the mean attribute weights of the Dirichlet model were obtained through bootstrapping. For the Dirichlet and MXL model, the fitted distribution of the attribute weights was also visually compared with the actual distribution of the attribute weights in the source data.

To assess the convergence of the MNL and Dirichlet models to the previously fitted population models $\hat{f}_{MNL}(\tilde{w})$ and $\hat{f}_D(w)$, we conducted a series of computational experiments with a varying number of respondents. The DCE data sets for the MNL models were constructed by simulating 6 discrete choice questions

from the set of 16 possible questions for each participant. The questions were resampled for each simulated respondent to minimize errors due to inefficient experimental design. The choice probabilities for the choice alternatives in these questions were obtained directly from the logit probabilities evaluated at $\hat{f}_{MNL}(\tilde{w})$. For the fitting of the Dirichlet models, attribute weights were randomly sampled from $\hat{f}_D(w)$. We varied the number of simulated respondents between 20 and 500 and repeated each simulation 100 times to assess the variance of the results. We measured convergence by calculating the Euclidean distance between the mean normalized attribute weights from the fitted models and the mean normalized attribute weights from $\hat{f}_{MNL}(\tilde{w})$ and $\hat{f}_D(w)$.

By measuring convergence to the (normalized) means of the previously fitted population models rather than to the sample mean of the attribute weights in the source data, we are able to assess convergence under ideal circumstances, where no bias is caused by misspecification of the preference model. In addition to the Euclidean

distance, we evaluated the MNL model coefficients' P values to understand when the hypothetical analyst could consider the results to be sufficiently accurate. Finally, we measured maximum acceptable risks of the adverse event (AE) attributes to assess whether the results in terms of key behavioral outputs are different from the results of individual model parameters.

All simulations were implemented in R. The MXL model was estimated using 5000 Halton draws. All program code and the full-source data set are available online.¹⁹

This research has received no external funding.

Results

The fitted Dirichlet, MNL, and MXL population models as well as the sample mean of the attribute weights in the source data are presented in Table 1. All models approximated the sample mean well, and the MNL and MXL models had reasonably good fits (adjusted $R^2 = 0.12$ and 0.13).

Figure 1 presents the results from the computational experiments assessing model convergence. For both the Dirichlet and MNL models, the estimated mean normalized attribute weights converged toward the population mean values in Table 1. The Dirichlet model seems to converge better than the MNL model: the mean attribute weights of the fitted Dirichlet models converged to within 0.05 distance of the mean of $\hat{f}_D(w)$ with a sample size of 100 in 97% of the simulations, whereas the fitted MNL models required a sample size of 240 to converge to the same distance of the normalized mean attribute weights of $\hat{f}_{MNL}(\bar{w})$ in 96% of the simulations. Similar results were observed when convergence was assessed in terms of maximum acceptable risks (see the results in the Supplementary Material).

Figure 2 presents convergence of the MNL model with respect to the P value of the least important attribute (moderate AEs). The results indicate that the MNL model consistently produced significant ($P < 0.05$) estimates with a sample size of 100 or higher. With sample sizes of 60 to 80, there were some simulations in which the analyst would not be able to conclude the significance of the estimate.

Figure 3 compares the distribution of the attribute weights in the original study with the distribution of the attribute weights for the fitted Dirichlet and MXL models. Samples drawn from the fitted MXL and Dirichlet models have similar spread over the preference space, although the Dirichlet model seems to have a slightly

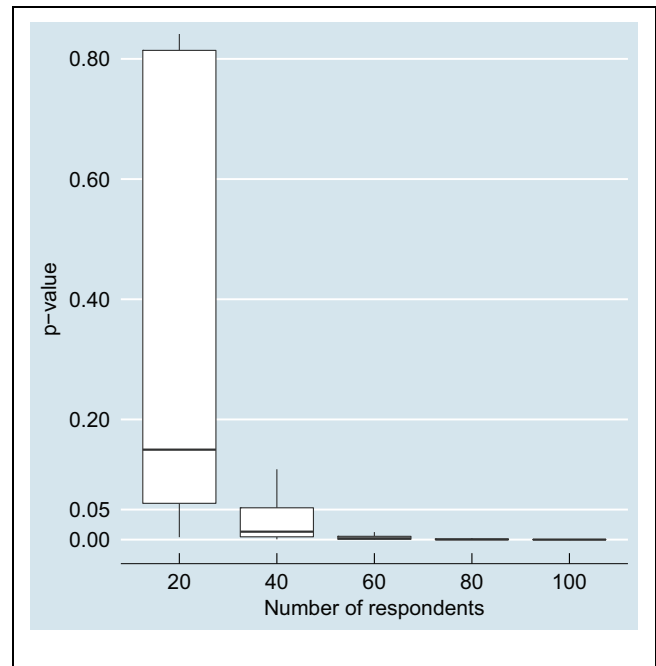


Figure 2 Significance (P value) of the least-important attribute (moderate adverse effects) in the multinomial logit model, with sample size varying from 20 to 100; the P value was <0.05 in all simulations in which the number of respondents was >100 .

higher dispersion than the MXL model. Both models seem to describe the source data reasonably well.

Discussion

Our computational experiments demonstrated that the Dirichlet model is likely to have smaller sample size requirements than the MNL model in modeling population benefit-risk preferences. Although we have no quantitative evidence of differences between the Dirichlet and MXL approaches, the full-sample preference distributions seemed similar, which indicates that the Dirichlet distribution may also be appropriate for modeling preference heterogeneity in benefit-risk tradeoffs. Importantly, our results indicate that the Dirichlet distribution is able to represent the population benefit-risk tradeoffs once they are captured using an elicitation technique, such as the choice-based matching that was applied in the source data study. This implication has direct practical relevance for treatment benefit-risk analyses using methods that apply the Dirichlet distribution^{20,21}: our results demonstrate that the full distribution, including the concentration parameter that has previously been undefined,

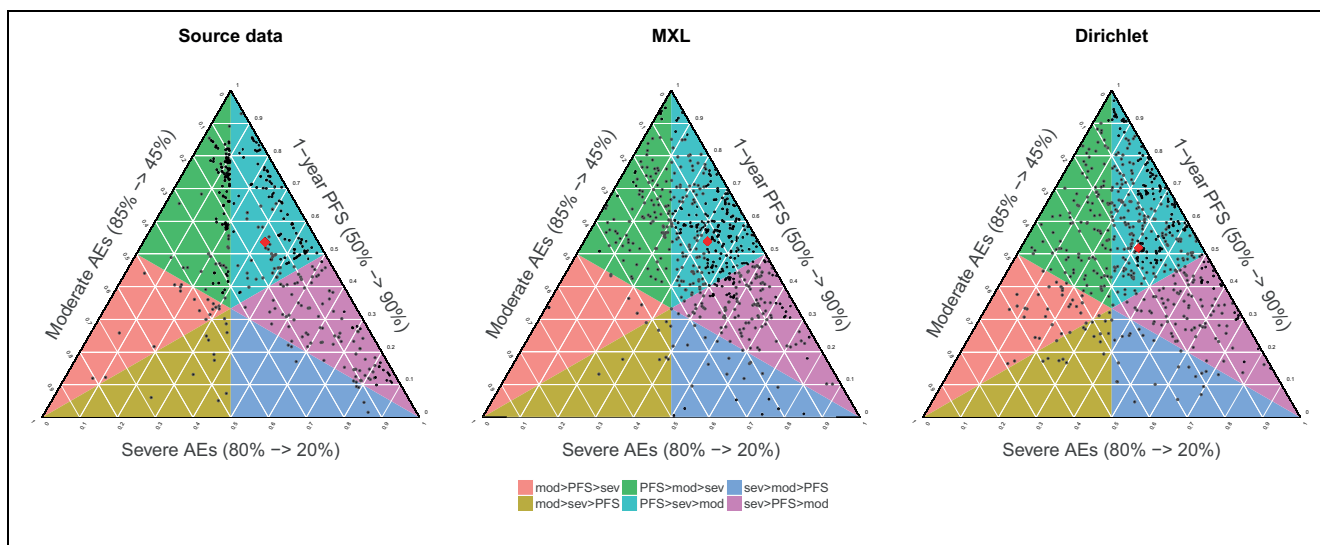


Figure 3 Weights from the original study (left) and the same number of samples ($N = 560$) from the MXL model (center) and Dirichlet model (right); red dots indicate sample mean (source data) and distribution means (MXL and Dirichlet). AE, adverse event; MXL, mixed logit; PFS, progression-free survival.

can reliably be estimated with reasonably small sample sizes.

Fitting a Dirichlet distribution with a standard maximum likelihood procedure requires per-respondent tradeoff weights to be available in complete format. These are usually obtained with a direct elicitation procedure, which is generally thought to be more demanding to complete than indirect procedures such as DCEs, as they require preferences to be expressed in precise cardinal terms. Therefore, direct elicitation procedures often require facilitation, making their application a resource-intensive exercise with larger samples.²² However, once the per-respondent preferences are available, understanding their distribution requires less modeling than what is needed to analyze discrete choice data.

This study has some important limitations. First, we conducted experiments on only a single data set that contained 3 attributes. Most health preference studies are conducted on larger sets of benefit, risk, and process attributes. However, the Dirichlet distribution is well understood, and we would not expect the estimate precision to suffer more from an increase in dimensionality than the MNL model. Furthermore, using only a three-attribute data set has the additional advantage of the preference space being 2 dimensional, and therefore, it can be easily visualized. Future research should assess the Dirichlet approach in studies with more attributes. Second, we did not compare the Dirichlet and MXL models in the experiments because 1) MXL estimation is

much more time-consuming than MNL estimation and 2) specifying preferences that adhere to the MXL distributional assumptions would have added an extra layer of complexity to the experiments. Third, we considered the Dirichlet model only for the case in which respondent preferences are available in a complete format. In practice, there may be partial or incomplete preference data for some respondents, such as ranking of the attribute scale swings instead of the exact tradeoff weights. Future research should consider estimation and convergence of the Dirichlet model in such cases.


Acknowledgments

We thank Sebastian Heidenreich (Evidera) for his helpful suggestions on simulating MNL responses and Vibha Shukla (Evidera) for editorial support.

Disclaimer

The views presented here are those of the authors and not of their organizations.

ORCID iD

Tommi Tervonen  <https://orcid.org/0000-0001-7303-500X>

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* Web site at <http://journals.sagepub.com/home/mdm>.

References

1. Egbrink MO, IJzerman M. The value of quantitative patient preferences in regulatory benefit-risk assessment. *J Mark Access Health Policy*. 2014;2.
2. Eichler HG, Abadie E, Baker M, Rasi G. Fifty years after thalidomide; what role for drug regulators? *Br J Clin Pharmacol*. 2012;74(5):731–3.
3. US Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Center for Biologics Evaluation and Research. 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. Available from: <https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM446680.pdf>
4. Hauber AB, Fairchild AO, Johnson FR. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. *Appl Health Econ Health Policy*. 2013;11(4):319–29.
5. Johnson FR, Lancsar E, Marshall D, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value Health*. 2013;16(1):3–13.
6. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. *Pharmacoeconomics*. 2008;26(8):661–77.
7. 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.
8. Hauber AB, Gonzalez JM, Groothuis-Oudshoorn CG, 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):300–15.
9. Louviere JJ, Street D, Burgess L, Wasi N, Islam T, Marley AAJ. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *J Choice Model*. 2008;1(1):128–64.
10. Fischer GW, Jia J, Luce MF. Attribute conflict and preference uncertainty: the RandMAU model. *Manage Sci*. 2000;46(5):669–84.
11. Jessop A. Using imprecise estimates for weights. *J Oper Res Soc*. 2011;62(6):1048–55.
12. Moskowitz H, Tang J, Lam P. Distribution of aggregate utility using stochastic elements of additive multiattribute utility models. *Decis Sci*. 2000;31(2):327–60.
13. Postmus D, Richard S, Bere N, 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):44–51.
14. Johnson FR, Kanninen B, Bingham M, Özdemir S. Experimental design for stated-choice studies. In: Kanninen BJ, ed. *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Approach to Theory and Practice*. Dordrecht, the Netherlands: Springer; 2007. p 159–202.
15. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1–10.
16. Lesaffre E, Albert A. Partial separation in logistic discrimination. *J R Stat Soc Series B Methodol*. 1989;51.
17. Santner TJ, Duffy DE. A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1986;73(3):755–8.
18. Oehlert GW. A note on the delta method. *Am Stat*. 1992;46(1):27–9.
19. Tervonen T, Postmus D. tommite/pub-dirichlet-mnl: Code for “From individual to population preferences: comparison of discrete choice and Dirichlet models for treatment benefit-risk trade-offs” (Version v1-pub). Zenodo; 2019.
20. Saint-Hilary G, Cadour S, Robert V, Gasparini M. A simple way to unify multicriteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a Dirichlet distribution in benefit-risk assessment. *Biom J*. 2017;59(3):567–78.
21. Li K, Yuan SS, Wang W, et al. Periodic benefit-risk assessment using Bayesian stochastic multi-criteria acceptability analysis. *Contemp Clin Trials*. 2018;67:100–8.
22. Tervonen T, Gelhorn H, Sri Bhashyam S, 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):1483–91.