Title

An investigation of the stability of patients’ treatment preferences over the course of a clinical trial.

Running head

Treatment preference stability in a clinical trial

Authors

Paul F. Allanson, PhD
University of Dundee School of Business, University of Dundee
Dundee, United Kingdom

Eric A. Brown, PhD
University of Dundee School of Business, University of Dundee
Dundee, United Kingdom

Daniel Kopasker, PhD
Health Economics Research Unit, University of Aberdeen
Aberdeen, United Kingdom

Andrzej Kwiatkowski, PhD
University of Dundee School of Business, University of Dundee
Dundee, United Kingdom

Contact

Professor Paul Allanson
University of Dundee School of Business
3 Perth Road
Dundee DD1 4HN
United Kingdom
p.f.allanson@dundee.ac.uk
Phone: +44 01382 386164
Fax: +44 01382 384691.

Funding

The study was funded by the National Institute for Health Research (NIHR) Research for Patient Benefit programme (PB-PG-010-21244).

Financial disclosure

None reported.

Precis

This is the first study to examine the temporal stability of patients’ treatment preferences as an integral component of a clinical trial.

Acknowledgements

Catherine Harwood, PhD as chief investigator and Charlotte Proby, MA, John Lear, MD, Adele Green, PhD, and Keith Wheatley, DPhil as co-investigators were responsible for the original concept and design of the main clinical trial and, together with other members of the trial team, supervised the collection of the data for this sub-study. The Cancer Research UK Clinical Trials Unit provided trial coordination. We thank the anonymous reviewers for their valuable comments.

Source of financial support: The study was funded by the National Institute for Health Research (NIHR) Research for Patient Benefit programme (PB-PG-010-21244).

Word count

3676

Number of pages

28

Number of figures	2
Number of Tables	4
Appendix	Pages: 2. Figures: 0. Tables 2.

Abstract

Objectives

The usefulness of DCEs to inform clinical guidelines rests on the assumption that patients facing the same treatment choice at different points in time will express the same preferences. This study provides the first investigation to our knowledge to specifically focus on the stability of patients' treatment preferences over the course of a clinical trial.

Methods

The same Decision Choice Experiment (DCE) was completed by participants at baseline and final post-treatment assessment in a trial of the efficacy of alternative topical treatments for actinic keratosis as a means for the prevention of skin cancer. The study assesses both the consistency of stated treatment choices and the stability of population-level preference parameter estimates, and analyses how the former is influenced by design aspects of the DCE.

Results

No evidence is found of population-level preference parameter instability over the course of the trial despite only a moderate strength of choice consistency. Choice consistency is negatively related to task difficulty with weak evidence of the existence of ordering effects over the sequence of choice tasks.

Conclusions

The results provide no evidence that the timing of a DCE within a clinical trial significantly influences population-level treatment preference estimates.

Highlights

i. What is already known about the topic?

Patients’ treatment preferences are increasingly being elicited using Discrete Choice Experiments (DCEs) but there is only limited evidence available on the stability of such preferences over the course of a treatment.

ii. What does the paper add to existing knowledge?

This study is the first to specifically focus on the stability of patients’ treatment preferences over the course of a clinical trial. No evidence is found that population-level preference parameters change between the baseline and final post-treatment assessments despite only a moderate strength of choice consistency. The design of the DCE is shown to affect choice consistency.

iii. What insights does the paper provide for informing healthcare-related decision making?

The routine incorporation of DCEs into clinical trials would generate information on patient treatment preferences to complement existing cost effectiveness analyses in helping shape clinical guidelines. The findings of our study provide no evidence that the timing of the DCE within a trial significantly influences population-level treatment preference estimates.

Introduction

Discrete Choice Experiments (DCEs) are increasingly being used to elicit patients' treatment preferences in order to take account of healthcare attributes beyond health outcomes.[1] The usefulness of such DCEs to inform clinical guidelines rests on the assumption that patients facing the same healthcare choice at different points in time will express the same preferences, making the temporal stability of preferences an important concern. This issue has been addressed in a number of previous studies in a variety of decision-making contexts, producing mixed results in terms of both the consistency of individual choices and the stability of population-level preference parameter estimates over time.[2,3] Respondents presented with the same choice set on a second occasion are found to choose the same option more often than might be expected by chance, but reported levels of agreement in individual choices between repeated surveys range from below 60% in Liebe et al. and Schaafsma et al. to over 80% in Ryan et al. and Gamper et al.[4-7] The first two studies both reject the stability of preference weights whereas the latter two are unable to do so.

This paper adds to the existing healthcare literature by providing a study of the temporal stability of actinic keratosis (AK) patients' preferences for topical treatments over the course of a clinical trial that investigates the efficacy of alternative skin cancer prevention regimens. In an early contribution, San Miguel et al. find that preferences for out-of-hours health care for children do not change with service experiences.[8] However, the closest comparable study is Skjoldborg et al. which explores rheumatoid arthritis patients' treatment preferences using a DCE that was administered three times to the same group of patients with 4 months between each survey to minimise possible memory effects.[9] Reported choice consistency rates are 76% between the first two surveys, and 87% between the last two in the subset of individuals who made identical choices in the first two surveys. Choice model estimates of attribute preference weights are not found to differ significantly between the surveys. The main aim of this study is to address the open question whether patients' treatment preferences change between the start and end of a clinical trial where this might occur as a result of their participation in the trial.

Özdemir et al. (2010) provides a meta-analysis of a number of healthcare DCE studies which explores the factors affecting the frequency with which respondents make identical choices when presented with the same choice set twice within a single survey.[10] A secondary goal of the study is to explore the determinants of choice consistency between repeated administrations of the same survey, which has not previously been done within a healthcare setting. We focus on the difficulty of individual choice tasks as measured by either the entropy of the predicted choice probabilities,[4] which provides an information theoretic measure of task complexity,[11] or the difference in predicted utility between the best two choices.[12] We also consider the possible impact of the ordering of choice tasks on consistency, perhaps due to learning or fatigue effects.[13]

Methods

The research was conducted as part of Squamous cell carcinoma Prevention in Organ transplant recipients using Topical treatments (SPOT), a multi-centre, randomised, 3-arm feasibility study comparing topical treatments of AK as a strategy for prevention of skin cancer.[14] SPOT was approved by the Research Ethics Committees of the participating sites and all participants provided written informed consent.

The same DCE was administered twice in SPOT, the first time as part of the initial baseline assessment to all study participants and again, 15 months later, as part of the final post treatment assessment to the subset of participants entered into the clinical trial. Kopasker et al. discuss the design, piloting and administration of the DCE, which formed part of a written questionnaire completed in clinic with a research nurse in attendance to check respondents' understanding, and present preference weight estimates based on the baseline assessment.[15] The sample size was determined by a power calculation for the main SPOT study, not the DCE, but the Johnson and Orme rule of thumb was used to check that this was adequate to detect the main effects in the choice model analysis.[16] The procedure in the retest assessment was the same as for the baseline test.

Participants were presented with a series of choice sets, each consisting of two AK treatment alternatives (A and B) with different hypothetical combinations of attribute levels, and a 'no treatment' opt-out option. Three attributes relate to the *burden* of medication (treatment regimen, severity of local skin reaction, and occurrence of systemic side effects) and two to the *efficacy* of treatment (improvement in skin appearance and reduction in skin cancer risk). Attribute levels were comparable to those of currently prescribed creams to ensure clinical relevance, with three levels specified for skin cancer risk reduction and two levels each for the four other attributes.

The DCE employed a D-efficient orthogonal main effects plan[17] consisting of 12 of the 48 possible combinations of treatment attribute levels. To validate participant responses, a further two choice sets were added to the DCE: the first checked for rationality by specifying one treatment option with unambiguously higher medical burden and lower clinical efficacy; the second checked for choice consistency by repeating one of the main choice sets but with the labelling of treatments A and B reversed, where the same choice set was repeated in all questionnaires. Importantly, the sequence of choice sets within the DCE was randomly generated so respondents did not face them in the same order either as each other or in the two assessments.

Statistical Analysis

Choice consistency is measured as the proportion of cases in which respondents make the same choice of best option when faced with the same choice set, either within the same DCE or in the retest. Cohen's kappa statistic $\kappa = \left(p_o - p_e \right) / \left(1 - p_e \right)$, where p_o is the observed agreement rate and p_e is the probability of chance agreement, is also calculated to provide an estimate of agreement that is corrected for chance.[18] Lack of choice consistency within a DCE may arise either because choice behaviour is intrinsically stochastic, as is assumed to be the case in random utility theory, or due to decision errors resulting from factors such as the limited ability of respondents to discriminate between options in a choice set, inattention or fatigue. Inconsistencies between repeated DCEs might additionally reflect shifts in preferences. In our study, preference changes for treatment attributes may plausibly result from the knowledge and experience gained by patients through their participation in the trial.

Preference stability is inferred from the equality of choice model parameter estimates obtained from the two assessments, with identification contingent on model specification. We make use of a basic multinomial logit (MNL) model incorporating a single alternative-specific constant to capture the value of treatment *per se*. We do not consider more general random parameter logit specifications because of the relatively small sample sizes. Thus, the utility that participant *i* assigns to treatment *j* in choice set *s* at assessment *t*, U_{ijst}, is given by:

$$U_{ijst} = \mu\beta_{0t}D_{js} + \sum_{k=1}^{K}\mu_t\beta_{kt}A_{jks} + \varepsilon_{ijst};\ i=1,\dots n;\ \ j=1,\dots J;\ s=1,\dots S;\ t=1,2 \qquad (1)$$

where $D_{js} = 1$ for $j \in \{A, B\}$ and zero otherwise such that, at assessment *t*, β_{0t} captures the preference for treatment; the population-level preference parameters β_{kt} (*k*=1,…*K*) indicate the marginal utilities of the treatment attributes A_k included in the DCE; and the set of error terms ε_{ijst} are independent and identically Gumbel distributed with standard deviation inversely related to the scale parameter μ_t.

Allowing for the possibility of scale heterogeneity between assessments is important as differences in scale can confound identification of differences in preference parameter estimates.[19,20] To investigate the stability of preferences we first estimate separate MNL models for the test and retest samples and a heteroscedastic conditional logit (HCL) model for the pooled sample in which the scale parameter is allowed to differ between the two assessments. A likelihood ratio test is performed to see if restricting the preference parameters to be equal results in a significant deterioration in model fit. As the null hypothesis cannot be rejected we also estimate a MNL model for the pooled sample and use these additional results to perform a likelihood ratio test of the equality of the scale parameter.

Finally, the determinants of choice consistency are analysed via the estimation of a fixed effects probit model:

$$P(D_{is} = 1 \mid x_{1is}, \dots x_{Vis}) = \Phi\left(\gamma_{0i} + \sum_{v=1}^{V}\gamma_v x_{vis}\right); i=1,\dots n; s=1,\dots S\ ; \qquad (2)$$

where variable x_{vis} ($v=1,\dots V$) is hypothesised to explain whether the option chosen by participant *i* in choice set *s* was the same in both assessments ($D_{is}=1$) or not $(D_{is}=0)$. Choice consistency is modelled as a function of choice task difficulty and choice set positioning. Our preferred measure of the former is given by the entropy of the predicted choice probabilities: $ENTROPY_s=-\sum_{j=1}^{J}P(j)\log P(j)\geq 0$, where $P(j)$ is the pooled MNL prediction of the probability of option *j* being chosen in choice set *s*.[3,21] We also consider the expected utility difference between the best and second best choices as a measure of choice task difficulty,[12] where these choices are always the two treatment options given the pooled MNL estimates: $UDIFF_s=\left|U_{As}-U_{Bs}\right|$ where U_{js} is the pooled MNL prediction of the utility of option *j* in choice set *s*. Average choice set position over the two assessments is given by $POSAV_{is}=\left(P1_{is}+P2_{is}\right)/2$, where $P1_{is}$ and $P2_{is}$ are respectively the test and retest positions of choice set *s* for individual *i*, which will vary across respondents due to the randomisation procedure. $POSAV_{is}$ will fully capture the influence of choice set position if the marginal impact of position on the value of the index function in (2) is constant. To allow for the possibility of non-constant marginal ordering effects we also include a position deviation variable $POSDEV_{is}=\sqrt{\left(P1_{is}{}^{2}+P2_{is}{}^{2}\right)/2}-POSAV_{is}$, which takes a minimum value of zero only if a choice set appears in the same position in both DCEs and is a decreasing function of average position for any given absolute difference in position between the two DCEs. The fixed effects specification fully controls for all time-invariant patient-specific factors.[22] We employ an estimator that incorporates an analytical correction for the bias due to the incidental parameter problem.[23]

Results

Sample characteristics

SPOT recruited 109 AK patients: 49 organ transplant recipients (OTRs) to take part in the clinical trial and a further 60 immunocompetent patients (ICPs) who only completed the initial baseline assessment. This analysis is limited to the OTR subsample of whom one patient, wrongly identified as an ICP in Kopasker et al.,[15] failed to answer the baseline DCE. Nine OTRs were subsequently not entered into the trial with no reason provided in four cases, withdrawal of consent in two cases, and one case each due to a requirement for other treatment, failure of screening test and deterioration in renal function. Accordingly, the final post treatment sample consisted of 40 OTRs, but only 35 completed the final assessment with the DCE left blank in four of these cases.

Supplementary Table 1 summarises patient characteristics at baseline. Patients were predominantly male with mean age 64. The majority had been most recently diagnosed with AK at least three years previously and considered their condition to be moderately serious in nature. Most had received prior treatments for AK with 55% reporting previous use of at least one topical treatment, specifically 5-fluorouracil cream, imiquimod cream or diclofenac gel. There is no evidence of differential attrition over the course of the trial, with no significant differences in baseline characteristics between those who did and did not provide retest DCE responses.

Classification of participants by DCE responses

Table 1 classifies participants by the pattern of their DCE responses in the two assessments. The proportions of respondents failing at least one of the validity tests were 17% and 10% in the test and retest DCEs respectively. More specifically, the corresponding proportions failing to make the same choice of best option in the repeat-choice task were 10% and 6%, yielding a weighted average agreement rate of 91% as a benchmark of within-test choice consistency. Of those passing both tests, 33% and 25% respectively in the two DCEs displayed ‘lexicographic’ choice behaviour – choosing the alternative that was the best with respect to one particular attribute in all 12 main choice sets – although only one respondent displayed such behaviour in both assessments. Following common practice,[24] respondents were excluded from the choice model analysis if they made either invalid or lexicographic choices, leaving those who passed both tests and revealed a willingness to trade gains in one attribute against losses in another at the levels specified in the DCE. These ‘traders’ accounted for 55% and 68% of DCE respondents respectively in the two assessments, with the 13 individuals who were willing to trade in both DCEs representing 27% of test and 42% of retest respondents.

Choice consistency

Table 2 presents choice consistency statistics. Choices are counted as consistent if a respondent made the same choice of best option in a choice set in both assessments, and inconsistent otherwise (including two instances where no choice was indicated in one of the assessments). The gross level of agreement is 73.1%, implying that respondents chose the same best option on average in 8.77 of the 12 main choice sets. This is higher than expected by chance, with the Cohen’s Kappa of 0.469 indicating a ‘moderate’ strength of agreement according to the commonly cited Landis and Koch benchmarks.[25] However, agreement rates differ markedly across respondents with the least consistent only making identical choices in 4 sets (i.e. 33% agreement) and the most consistent in 11 (92%). A subgroup analysis based on Fisher’s exact test revealed no significant association between agreement rates and a tripartite classification of respondents by DCE responses in the two assessments [$p = 0.936$]. Moreover, average agreement rates are not significantly lower among respondents who failed at least one validity test in at least one assessment compared to either those who made lexicographic choices in at least one assessment [t=−0.659, p=0.511] or traders in both assessments [t=−0.988, p=0.324].

Population-level preference parameter estimates

Table 3 presents choice model estimation results. The positive estimates of all attribute preference weights are consistent with *a priori* expectations, implying preferences for a lower medical burden and higher clinical efficacy given the attribute variable coding. The positive treatment specific constant implies that even the worst possible hypothetical treatment, with maximum burden and minimum efficacy, is preferable to no treatment. Comparing the pooled HCL model with the test and retest MNL models, the chi-square value is 6.361 (=–2*([–258.84] – ([–145.97] +[–109.69])), p=0.384 on 6df). Hence we are unable to reject the null of equality of marginal utilities between the test and retest samples. Furthermore, the estimated scale term in the HCL model is not significantly different from zero, implying that the error variance does not differ significantly between the two samples. This finding is confirmed by the chi-square value of 0.051 (=–2*([–258.86] – [–258.84]), p=0.823 on 1df) from a comparison of the pooled MNL and HCL models. Similar results are obtained if the choice model samples are expanded to include respondents with lexicographic choices and/or failed validity tests (see Supplementary Table 2).

Determinants of choice consistency

Choice consistency is modelled as a function of choice task difficulty and choice set positioning. Figure 1 shows that the agreement rate is negatively correlated with entropy: the probability of repeating a particular choice is higher, *ceteris paribus*, in a choice set with more dissimilar choice probabilities, and hence lower entropy, than in one where the probabilities are more even.

Figure 2 shows that there is a weak positive association between agreement rate and average choice set position.

Table 4 reports average partial effect estimates from the choice consistency model. We focus on the results for the sub-sample of respondents who passed both validity tests in both DCEs as these provide the stronger statistical evidence of the existence of ordering effects. Thus, a unit increase in entropy leads to a significant decrease of 0.596 in the probability of a consistent choice, implying a 33.3% (=−100*0.559*[−0.596]) difference in predicted agreement rates between the choice sets with the lowest and highest entropies. As expected, choice consistency increases as the options within a choice set become more dissimilar in terms of their probabilities of being chosen. The null of no ordering effects is rejected by a likelihood ratio test with a chi-square value of 6.016. The probability of agreement rises by 0.013 if a choice set is presented one position later on average, holding position deviation constant, with the simplest such case being when the choice set appears in the same position in both test and retest DCE both before and after the change. Noting that the complete DCE consists of 14 not 12 choice sets, the predicted agreement rate is 18.2% (=14*0.013) higher if a choice set appears last rather than first in both DCEs. There is also weak evidence of diminishing marginal ordering effects with a unit increase in position deviation leading to a 0.068 fall in the probability of a consistent choice holding average position constant, implying a maximum negative effect on the predicted agreement rate of 16.5% (=100*2.42*0.068 where 2.42=$\sqrt{(196+1)/2}-(14+1)/2$) when the choice set appears first in one DCE and last in the other. The use of the alternative measure of task difficulty does not lead to any substantive changes in the findings, with a 31.3% (=100*0.929*0.337) difference in predicted agreement rates between the choice sets with the largest and smallest utility difference.

Discussion

This study adds to the limited evidence currently available on the temporal stability of patients' treatment preferences. In particular, we provide the first investigation of this issue that has to our knowledge been conducted as an integral part of a clinical trial.

With regard to choice consistency, the strength of agreement between the options chosen by patients at the start and end of the clinical trial is only moderate. Some additional insight is provided by the direct comparison of agreement rates for the repeated choice set in the DCE, with the average within-test rate of 91% higher than the between-test rate of 80% for the same decision task. A one-sided hypothesis test provides no evidence of a lower degree of between-test than within-test consistency ($z = -1.397$; $p = 0.919$), such as might result *inter alia* from a shift in patients' treatment preferences.

Nevertheless, the inability to demonstrate evidence of preference parameter changes is perhaps surprising given contrary indications from other studies. In particular, Serra-Guillen et al. report that only 72% of AK patients treated in an RCT with one of the topical treatments tested in the SPOT trial would have been willing to repeat the treatment despite informed consent having been obtained from all participants.[26] More generally, studies of a range of healthcare services have found significant differences in preference patterns between groups with different levels of prior experience.[8,27,28] In our study most participants had received treatment for AK beforehand, with a majority having already received one or more topical treatments for the condition. It is therefore plausible that they maintained the same preferences as they were already reasonably well informed about the treatment attributes at the levels specified in the DCE.

Choice consistency is found to be negatively related to choice task difficulty, corroborating the findings of previous non-healthcare studies that have also shown large predicted differences in agreement rates between choice sets depending on task difficulty.[3] Intuitively, the harder the task the less likely that the same option will be chosen again when respondents are uncertain about their choices. Olsen et al argue that DCEs should contain a mix of tough and easy choice sets to ensure both the accuracy and precision of preference parameter estimates,[29] where the inclusion of some easy tasks should also help ensure their stability if preferences do not in fact change. We also obtain weak evidence of the existence of ordering effects, with the finding of positive but diminishing marginal effects consistent with the results of studies showing choice uncertainty to be greatest in the first few choice sets in a DCE,[13] possibly as the result of a positive but decreasing influence of learning through the sequence of choice sets.[29,13] Our randomisation of choice set position was specifically designed to mitigate against bias in preference parameter estimates due to learning or fatigue.

The study has a number of potential limitations. First the sample size may be too small to detect choice model parameter differences between the test and retest samples. However, further analysis shows the tests to be sufficiently powerful to detect significant differences in both preference and scale parameters between the similarly-sized OTR and ICP samples in the baseline assessment (see Supplementary Table 2): OTR patients may be expected to have a stronger preference for treatment given an approximately 100-fold increased risk of developing skin cancer.[31] Using the de Bekker-Grob et al sample size calculation[30] with initial beliefs about parameter values given by the pooled MNL estimates, both the test and retest sample sizes are sufficient to reliably detect preference differences between attribute levels at the 5% significance level. Given a larger sample it would have been of interest to formally test the impact of experience, if any, by treatment arm on preference stability. Second, the findings on population-level preference parameter stability may be sensitive to the choice model specification, but Kopasker et al. report robust estimates across a range of alternative logit specifications.[15] Finally, our findings are based on a single clinical trial of topical treatments for AK and may not be generalisable to trials of the same condition in other settings or of other conditions.

Conclusion

The routine incorporation of DCEs into clinical trials would generate information on patient treatment preferences to complement existing cost effectiveness analyses in helping shape clinical guidelines. This study is the first to examine the temporal stability of preferences within the context of a clinical trial, finding no evidence of population-level preference changes despite only a moderate degree of choice consistency. However, further studies are required to firmly establish the effect, if any, of DCE timing within a trial on preference parameter estimates, with prior treatment experience a potentially significant determinant of whether preferences are stable or not. The random assignment of participants to different treatment arms in an RCT provides an ideal setting to formally test for the influence of treatment experience on preference stability.

References

1. Ryan,M, Skåtun D, Major, K. Using Discrete Choice Experiments to Go Beyond Clinical Outcomes when Evaluating Clinical Practice. In Ryan M, Gerard K, Amaya-Amaya M, eds, Using discrete choice experiments to value health and health care. Dordrecht: Springer Netherlands, 2008.
2. Islam T, Louviere J. The stability of aggregate-level preferences in longitudinal discrete choice experiments. In Louviere J, Flynn T, Marley A eds, Best-Worst Scaling: Theory, Methods and Applications. Cambridge, England: Cambridge University Press, 2015.
3. Rigby D, Burton M, Pluske J. Preference stability and choice consistency in discrete choice experiments. Environ Resour Econ 2016; 65:441-461.
4. Liebe U, Meyerhoff J, Hartje V. Test-retest reliability of choice experiments in environmental valuation. Environ Resour Econ 2012; 53:389–407.
5. Schaafsma M, Brouwer R, Liekens I, de Nocker L. Temporal stability of preferences and willingness to pay for natural areas in choice experiments: a test-retest. Resour Energy Econ 2014; 38:243–260.
6. Ryan M, Netten A, Skatun D, Smith P. Using discrete choice experiments to estimate a preference-based measure of outcome - an application to social care for older people. J Health Econ 2006; 25:927–944.
7. Gamper E-M, Holzner B, King MT, et al. Test-retest reliability of discrete choice experiment for valuations of QLU-C10D health states. Value Health 2018; 21:958-966.
8. San Miguel F, Ryan M, Scott, A. Are preferences stable? The case of health care. J Econ Behav Organ 2002; 48:1–14.

9. Skjoldborg US, Lauridsen J, Junker P. Reliability of the discrete choice experiment at the input and output level in patients with rheumatoid arthritis. Value Health 2009; 12:153-158.

10. Ozdemir S, Mohamed AF, Johnson FR, Hauber AB. Who pays attention in stated-choice surveys? Health Econ 2010; 19:111–118.

11. Swait J, Adamowicz W. The influence of task complexity on consumer choice: a latent class model of decision strategy switching. J Consum Resour 2001; 28:135-148.

12. Morkbak RM, Olsen SB. A within-sample investigation of test-retest reliability in choice experiment surveys with real economic incentives. J Agric Resour Econ 2015; 59:375-392.

13. Carlsson F, Morkbak MR, Olsen SB. The first time is the hardest: A test of ordering effects in choice experiments. J Choice Model 2012; 5:19-37.

14. Harwood CA, Proby C, Lear J, et al. SPOT trial protocol (Version 1.0, June 10 2013). Available from: https://www.clinicaltrialsregister.eu/ctr-search/trial/2013-000893-32/GB. EudraCT Number: 2013-000893-32. Accessed July 10, 2019.

15. Kopasker D, Kwiatkowski A, Martin RN, et al. Patient preferences for topical treatment of actinic keratoses: a discrete choice experiment. Br J Dermatol 2018*;* 180:902-909.

16. Johnson R, Orme B. Getting the most from CBC. Sequim: Sawtooth Software Research Paper Series, Sawtooth Software Inc, 2003.

17. Street DJ, Burgess L, Viney R, Louvriere J. Designing discrete choice experiments for health care. In Ryan M, Gerard K, Amaya-Amaya M, eds, Using discrete choice experiments to value health and health care. Dordrecht: Springer Netherlands, 2008.

18. Cohen, J (1960). A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37–46.

19. Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. J Mark Res 1993; 30:305-314.

20. Vass CM, Wright S, Burton M, Payne K. Scale heterogeneity in healthcare discrete choice experiments: a primer. Patient 2018; 11:167-173.

21. Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Rev Pharmacoecon Outcomes Res 2017; 17:531-542.

22. Allison PD. Fixed effects regression models (Quantitative Applications in the Social Sciences, Book 160). Thousand Oaks, CA: Sage Publications Inc, 2009.

23. Fernandez-Val I, Weidner M. Individual and time effects in nonlinear panel models with large N, T. J Econom 2016; 192:291-312.

24. Hess S, Rose JM, Polak JW. Non-trading, lexicographic and inconsistent behaviour in stated choice data. Transport Res Part D 2010; 15:405-417.

25. Landis JR, Koch CG. The measurement of observed agreement for categorical data. Biometrics 1977; 33:159-174.

26. Serra-Guillen C, Nagore E, Hueso L, et al. A randomized comparative study of tolerance and satisfaction in the treatment of actinic keratosis of the face and scalp between 5% imiquimod cream and photodynamic therapy with methyl aminolaevulinate. Br J Dermatol 2011; 164:429–433.

27. Ryan M, Ubach C. Testing for an experience endowment effect within choice experiments. Appl Econ Lett 2003; 10:407–410.

28. Neuman T, Neuman E, Neuman S. Explorations of the effect of experience on preferences for a health-care service. J Socio Econ 2010; 39:407–419.

29. Olsen SB, Lundhede TH, Jacobsen JB, Thorsen BJ. Tough and easy choices: testing the influence of utility difference on stated certainty-in-choice in choice experiments. Environ Resour Econ 2011; 49:491–510.

30. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient 2015; 8:373–384.

31. Harwood CA, Mesher D, McGregor JM, et al. A surveillance model for skin cancer in organ transplant recipients: A 22-year prospective study in an ethnically diverse population. Am J Transplantation 2013; 13:119-129.

Table 1 Classification of participants by DCE responses in baseline and final assessments

				Post-treatment final assessment (n=35)						
								Passed both tests		
			Dropped out of study	No responses	Failed rationality test only	Failed consistency test only	Failed both tests	Lexicographic choices	Trader	Total
Baseline assessment (n=49)		No responses	-	-	-	-	-	-	1	1
		Failed rationality test only	2	1	-	-	-	-	-	3
		Failed consistency test only	1	-	-	1	-	-	1	3
		Failed both tests	0	1	-	-	-	1	-	2
	Passed both tests	Lexicographic choices	5	-	-	1	-	1	6	13
		Trader	6	2	1	-	-	5	13	27
		Total	14	4	1	2	0	7	21	49

DCE indicates discrete choice experiment; Failed both tests, failure of both rationality and consistency tests; Failed consistency test only, failure to choose the same option in the repeated decision task; Failed rationality test only, choice of dominated option in the rationality test choice set; Lexicographic choices, choosing the option that was the best with respect to one particular attribute in all 12 main choice sets; trader, choices revealed a willingness to trade gains in one attribute against losses in another at the levels specified in the DCE.

Table 2. Main choice set agreement rates

		Subgroup analysis		
		Failed at least one test in at least one assessment	Passed both tests in both assessments	
	All respondents to both assessments		Lexicographic choices in at least one assessment	'Trader' in both assessments
Number of respondents	30	5	12	13
Number of choice sets	360	60	144	156
Agreement rate (%)	73.1	68.3	72.9	75.0
Cohen's Kappa	0.469	0.388	0.466	0.504
Number of choice sets per person	12	12	12	12
Agreement per person (mean)	8.77	8.20	8.75	9.00
Agreement per person (minimum)	4	4	6	5
Agreement per person (maximum)	11	11	11	11

DCE indicates discrete choice experiment; Failed at least one test in at least one assessment, failure of at least one of the rationality and consistency tests in at least one of the test and retest assessments; Lexicographic choices in at least one assessment, choosing the option that was the best with respect to one particular attribute in all 12 main choice sets in at least one assessment; trader in both assessments, choices revealed a willingness to trade gains in one attribute against losses in another at the levels specified in the DCE in both assessments.

Table 3. Choice model estimation results

	Test MNL	Retest MNL	Pooled HCL	Pooled MNL
Attributes				
Regimen: (reference level twice daily for 12 weeks)				
Daily for 1 week	0.726** (0.168)	0.366 (0.224)	0.561** (0.137)	0.573** (0.139)
Local skin reaction: (reference level severe)				
Mild	0.704** (0.167)	0.583** (0.188)	0.639** (0.133)	0.651** (0.124)
Systemic effects: (reference level flu-like symptoms)				
No other side effects	0.625** (0.170)	0.409* (0.190)	0.514** (0.131)	0.524** (0.128)
Skin appearance: (reference level moderate improvement)				
Big improvement	0.811** (0.181)	1.075** (0.214)	0.899** (0.148)	0.912** (0.137)
Cancer risk: (reference level 20% fall)				
50% fall	1.377** (0.329)	1.712** (0.287)	1.469** (0.256)	1.491** (0.220)
60% fall	1.873** (0.328)	1.946** (0.367)	1.842** (0.288)	1.873** (0.242)
Treatment specific constant	16.667** (0.418)	15.598** (0.459)	19.681** (1.039)	16.264** (0.309)
Logarithm of relative scale parameter between assessments			0.038 (0.215)	
Number of observations†	972	753	1725	1725
Number of respondents	27	21	48	48
Log-likelihood	-145.97	-109.69	-258.83	-258.86

HCL indicates heteroscedastic conditional logit; MNL, multinomial logit.

†One respondent failed to indicate a preferred option in one choice set in the retest assessment.

Robust standard errors (in round brackets). *p<0.05, **p<0.01.

Table 4. Panel fixed effects probit model of choice consistency

Average partial effects	Sub-sample passing both validity tests in both assessments		Sub-sample with responses in both assessments	
ENTROPY	-0.596^{**}	-	-0.517^{**}	-
	(0.130)		(0.115)	
	[0.000]		[0.000]	
UDIFF	-	0.337^{**}	-	0.309^{**}
		(0.080)		(0.073)
		[0.000]		[0.000]
POSAV	0.013	0.013	0.010	0.010
	(0.008)	(0.008)	(0.008)	(0.008)
	[0.116]	[0.113]	[0.202]	[0.210]
POSDEV	−0.068	−0.073	−0.055	−0.059
	(0.042)	(0.042)	(0.039)	(0.039)
	[0.102]	[0.082]	[0.160]	[0.131]
Number of observations	300	300	360	360
Number of respondents	25	25	30	30
Log-likelihood	−138.51	−141.58	−172.14	−174.12
Pseudo-R^2	0.194	0.177	0.179	0.170
χ^2 test of no ordering effects on 2df	6.016^{*}	6.426^{*}	4.238	4.500
	[0.049]	[0.040]	[0.120]	[0.105]

ENTROPY indicates choice set entropy; POSAV, average choice set position; POSDEV, deviation in choice set position; UDIFF, predicted utility difference.

Respondent fixed effects not reported.

Robust respondent-clustered standard errors (in round brackets). p-values in [square brackets].
* $p<0.05$, ** $p<0.01$.

Figure 1. Choice consistency by choice set entropy

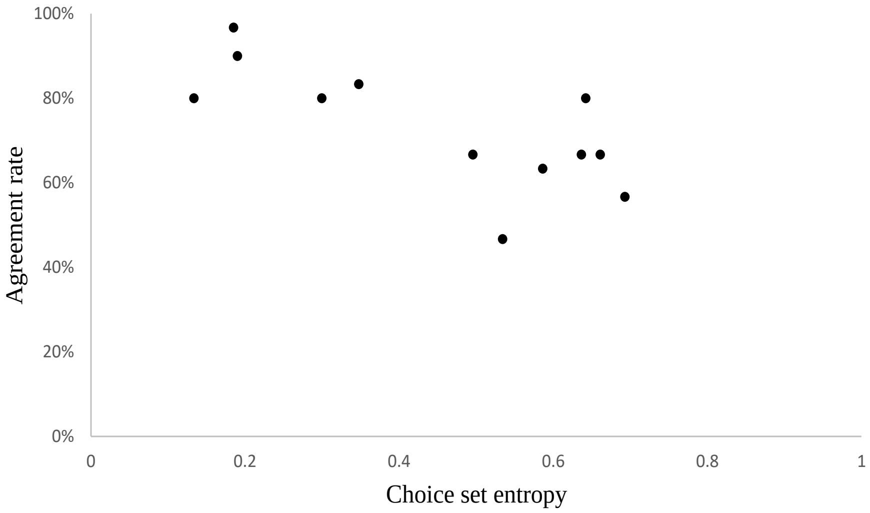
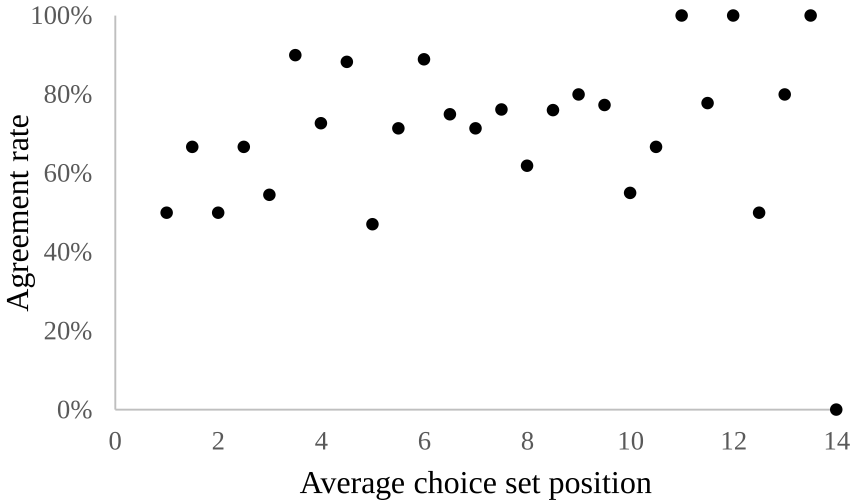


Figure 2. Choice consistency by average choice set position



Appendix

Supplementary Table 1

Patients' characteristics at baseline assessment

Patient Characteristic	Baseline assessment sample (n=49)	Retest DCE respondents (n=31)	Dropouts and non-respondents (n=18)	*p*-value
Age	64.40 [45.83-82.44]	63.87 [50.45-73.73]	65.38 [45.83-82.44]	0.542^{a}
Male	79.59% (39)	74.19% (23)	88.89% (16)	0.227^{a}
Age first left education				0.746^{b}
16 years or less	38.78% (19)	32.26% (10)	50.00% (9)	
17-19 years old	26.51% (13)	29.03% (9)	22.22% (4)	
20 years or over	32.65% (16)	35.48% (11)	27.78% (5)	
Not supplied	2.04% (1)	3.23% (1)	0.00% (0)	
Time since most recent diagnosis				0.069^{b}
Less than 1 day	2.04% (1)	0.00% (0)	5.56% (1)	
Less than 3 months	10.2% (5)	9.68% (3)	11.11% (2)	
Between 3 months and 1 year	6.12% (3)	0.00% (0)	16.67% (3)	
Between 1 and 3 years	0.00% (0)	0.00% (0)	0.00% (0)	
Between 3 and 10 years	36.73% (18)	35.48% (11)	38.89% (7)	
More than 10 years	42.86% (21)	51.61% (16)	27.78% (5)	
Not supplied	2.04% (1)	3.23% (1)	0.00% (0)	
Self-rated seriousness of AK condition				0.251^{b}
1. Not serious	2.08% (0)	3.23% (1)	5.88% (1)	
2	8.33% (4)	9.68% (3)	0.00% (0)	
3	31.25% (15)	35.48% (11)	23.53% (4)	
4. Moderately serious	27.08% (13)	22.58% (7)	35.29% (6)	
5	20.83% (10)	19.35% (6)	23.53% (4)	
6	4.17% (2)	0.00% (0)	11.76% (2)	
7. Very serious	6.25% (3)	9.68% (3)	0.00% (0)	
Previous treatment for AK	85.71% (42)	90.32% (28)	77.78% (14)	0.235^{a}
Past cryotherapy	67.35% (33)	70.97% (22)	61.11% (11)	0.489^{a}
Past photodynamic therapy	12.24% (6)	12.90% (4)	11.11% (2)	0.857^{a}
Past Diclofenac gel	4.08% (2)	6.45% (2)	0.00% (0)	0.281^{a}
Past skin surgery	38.78% (19)	29.03% (9)	55.56% (10)	0.069^{a}
Past 5-fluorouracil cream	46.94% (23)	54.84% (17)	33.33% (6)	0.152^{a}
Past imiquimod cream	26.53% (13)	29.03% (9)	22.22% (4)	0.612^{a}
Past other treatment	14.29% (7)	16.13% (5)	11.11% (2)	0.637^{a}
Past treatment not known	6.12% (3)	6.45% (2)	5.56% (1)	0.902^{a}
DCE difficulty (5=highest)	2.85 [1-4]	2.80 [1-4]	2.94 [1-4]	0.622^{a}

DCE indicates discrete choice experiment.

Continuous variables show mean and range [in square brackets]. Dummy variables show percentage and number (in round brackets). ^{a}t-test. ^{b} Fisher's exact test.

Supplementary Table 2

Choice model parameter stability tests for alternative sub-sample comparisons

			LR test of equality of:	
		χ^2	preference parameters (*df*=6)	scale parameter (*df*=1)
OTR Traders in test DCE (*n*=27)	*vs.*	OTR Traders in retest DCE (*n*=21)	6.361 [0.384]	0.051 [0.823]
All OTR passed both tests in test DCE (*n*=40)	*vs.*	All OTR passed both tests in retest DCE (*n*=28)	11.965 [0.063]	2.006 [0.156]
OTR traders & failed at least one test in DCE (*n*=35)	*vs.*	OTR traders & failed at least one test in DCE (*n*=24)	6.570 [0.362]	0.113 [0.737]
All OTR respondents in test DCE (*n*=48)	*vs.*	All OTR respondents in retest DCE (*n*=31)	7.002 [0.321]	1.624 [0.203]
OTR Traders in test DCE (*n*=27)	*vs.*	ICP Traders in test DCE (*n*=31)	15.194* [0.019]	6.366* [0.012]

DCE indicates discrete choice experiment; Failed at least one test, failure of at least one of the rationality and consistency tests; ICP, immunocompetent patients; LR, likelihood ratio; OTR, organ transplant recipients; Passed both tests, pass in both the validity tests; Trader, choices revealed a willingness to trade gains in one attribute against losses in another at the levels specified in the DCE.

p-values in [square brackets]. *p<0.05.