Duke Clinical Research Institute

From Thought Leadership to Clinical Practice

Quantifying Patient Preferences for Leadless Pacemaker Devices

Prepared for Abbott Laboratories

[redacted]

Duke Clinical Research Institute, Duke University

Statistical Analysis Plan

December 21, 2021

[redacted]

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Definition
AIC	Akaike information criterion
BIC	Bayesian information criterion
BLRT	Bootstrap likelihood ratio test
DCE	Discrete-choice experiment
DCRI	Duke Clinical Research Institute
[redacted]	[redacted]
FDA	Food and Drug Administration
IDE	Investigational Device Exemption
LCA	Latent-class analysis
MAR	Maximum-acceptable risk
[redacted]	[redacted]

1. STUDY DESIGN OVERVIEW

The primary objective of this study is to quantify patients' willingness to accept tradeoffs among pacemaker device features. To achieve the study objective, the [redacted] at Duke Clinical Research Institute (DCRI) has developed a discrete-choice experiment (DCE) survey instrument that adheres with guidance issued by Food and Drug Administration (FDA) in August 2016 and best practices identified by stated-preference experts (Bridges et al., 2011; Reed Johnson et al., 2013).

In the DCE, respondents are asked to select their preferred treatment options in a series of experimentally controlled pairs of hypothetical devices. The devices are defined by pacemaker type (pacemaker with leads, leadless pacemaker-removable, and leadless pacemaker- not removable), years of battery life (15 years, 12 years, 8 years, 5 years), years since government approval (10 years, 2 years), discomfort for 6 months (no discomfort, discomfort), and device-related risks of adverse events (risk of infection requiring removal and antibiotics and risk of complication requiring an operation). To gain insights about device features that are inextricably linked to a specific type of pacemaker that were not amenable to inclusion in a DCE, a ranking exercise was designed to elicit information on participants' relative concern for these device features (insertion procedure, where the device is implanted/resides in the body, the need for leads, whether the device is externally visible, and duration of activity restrictions after the device is implanted).

To prepare respondents for the DCE choice questions, the survey instrument includes carefully worded descriptions of all attributes, comprehension questions, a tutorial for probabilistic attributes, and simplified practice choice questions. Development and pretesting the survey instrument have been described in a separate memo.

Abbott will take primary responsibility for working with operational teams and Investigational Device Exemption (IDE) study sites to develop plans for recruitment and assist with their dissemination to study sites. DCRI will review site recruitment and training materials developed by Abbott for the patient preference study. Respondents eligible to participate in the online patient-preference study will be recruited from all adult patients who are referred to a study site for evaluation for a pacemaker or with a known indication for a de-novo pacemaker. Study sites will track which respondents agreed to be screened for the IDE trial and which respondents ultimately enrolled in the IDE trial.

This document provides details about the DCE and its experimental design, the ranking exercise, and the plan for statistical analysis.

1.1. Discrete-Choice Experiment

1.1.1. Attributes and Levels

At the core of every discrete-choice study is a set of treatment features or attributes that define the constructed treatments evaluated in the study. In a series of choice questions, respondents choose between two or more of these constructed treatments, where the levels of benefit and risk attributes are varied experimentally. The attributes and levels included in this study are shown in Table 1, and an example choice question is shown in Figure 1.

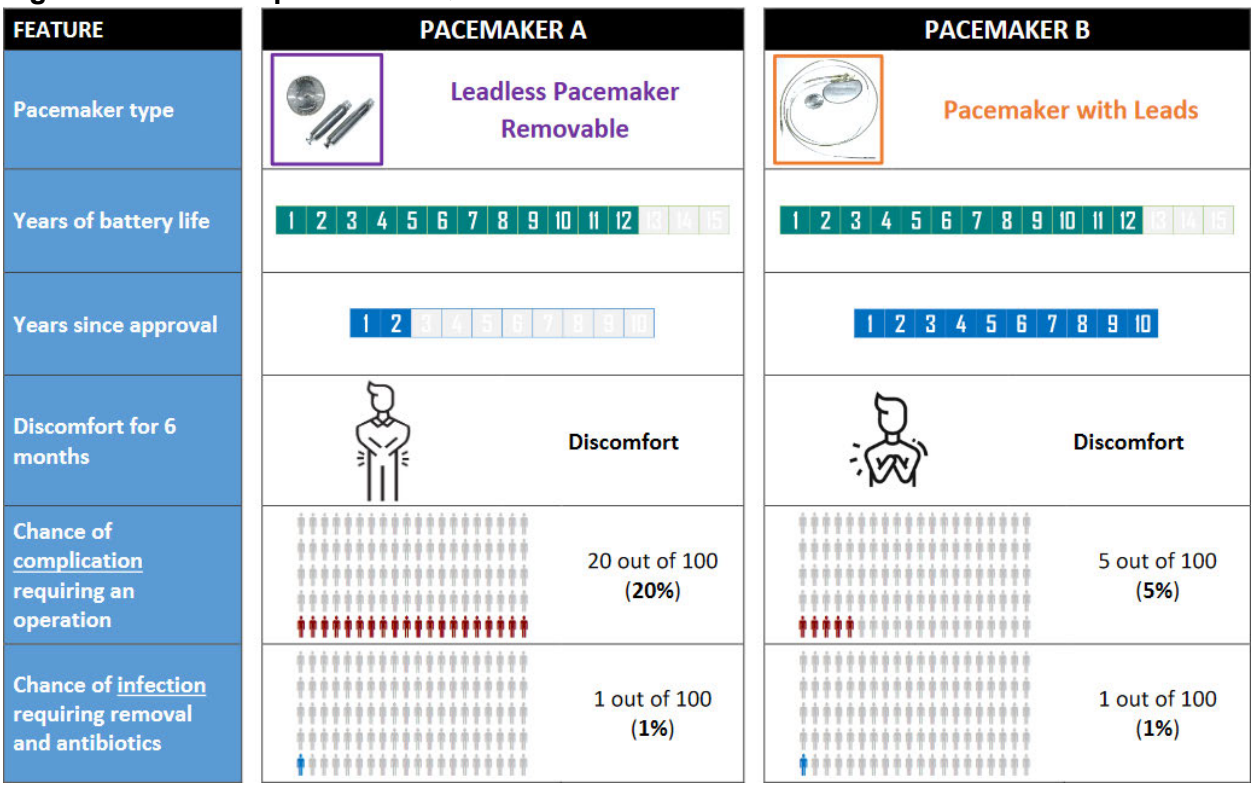
Table 1. Study Attributes and Levels

Attributes	Levels
Pacemaker type	• Pacemaker with leads • Leadless pacemaker removable • Leadless pacemaker not removable
Years of battery life	• 5 years • 8 years • 12 years • 15 years
Years since government approval	• 2 years • 10 years
Discomfort for 6 months	• No discomfort • Discomfort
Chance of complication requiring an operation	• 1% • 5% • 10% or 20%*
Risk of infection requiring removal and antibiotics	• 1% • 5% • 10% or 20%*

*Scope test – each respondent will be assigned to either 10% or 20% as the highest risk level for infection and complication risks.

As shown in Figure 1, each of the choice questions in this study consists of a choice between two pacemaker options in which attribute levels vary according to the experimental design.

Figure 1. Example Choice Question



1.1.2. Scope Test

As indicated in Table 1, each respondent will be randomized into a study arm in which 10% or 20% is shown as the highest risk level for chance of complication requiring an operation and chance of infection requiring removal and antibiotics. A scope test is designed to evaluate whether respondents are responsive to the magnitude of risk levels shown rather than rescaling the levels as ordinal measures (e.g. low, medium, high). Thus, we hypothesize that respondents randomized to the high-risk arm (i.e. 20% shown as the highest risk level) will have more negative preferences for the highest risk levels compared to respondents randomized to the low-risk arm (i.e. 10% shown as the highest risk levels). The scope test also allows us to evaluate whether the highest risk levels shown differentially impact preferences for the lower risk levels (i.e. 1% and 5%) shown in the low-risk and high-risk arms. We hypothesize that respondents' preferences in the low-risk and high-risk arms will not differ for the 1% and 5% risk levels for either adverse-event attribute.

1.1.3. Experimental Design

The experimental design determines the combinations of attribute levels that define each hypothetical pacemaker profile and the pairs of hypothetical profiles that will populate the choice questions in the DCE section of the survey.

The experimental design algorithms are designed to maximize the statistical efficiency (D-optimality[1]) of the choice questions. The efficiency of an experimental design is partly related to level balance—the number of times that each level and pair of levels is shown in the design.

By identifying potential designs based on D-optimality, we can control the statistical properties of the complete design. However, blocking, incomplete responses, and potential cognitive factors (e.g., unaccounted interaction effects, simplifying heuristics, etc.) can affect our ability to identify all preference parameters.

The smallest feasible experimental design requires as many questions as preference parameters to be estimated (i.e. total number of levels less total number of attributes). Accounting for the number of attribute levels (two 2-level attributes, three 3-level attributes, and one 4-level attribute) and assuming that all attribute-level variables will be effect-coded, this implies a design with at least 11 choice questions. However, to cover more unique combinations of attribute levels (Janssen, Hauber, & Bridges, 2018) within and across pacemaker alternatives, we used algorithms by Kuhfeld in SAS (Kuhfeld, 2010) to generate to generate orthogonal experimental designs from the full factorial design of 768 profile pairs. Efficient designs suggested by the algorithms included 48 and 96 profile pairs.

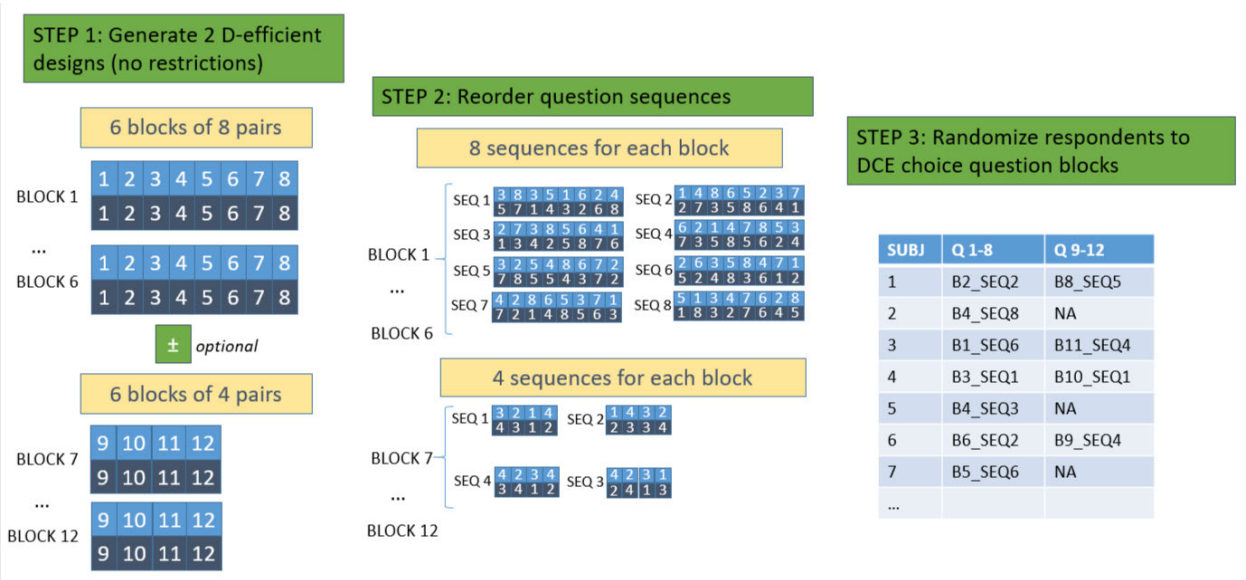
Our pretest interviews demonstrated that respondents were able to assess at least 8 choice questions without becoming overly fatigued. Considering the expected sample size of 300 and the plan to generate multiple sequences of choice questions within blocks for implementation in Sawtooth, we chose to generate a design with 48 profile pairs, split into 6 blocks of 8 choice questions. We generated several designs with 48 pairs of hypothetical pacemakers. No restrictions were implemented in the design, thus all levels across attributes could be combined to represent hypothetical pacemaker alternatives. The resulting designs had very similar statistical efficiency. DCRI chose a design with acceptable balance between levels.

[1] D-optimality is a criterion that supports the maximization D-efficiency in an experimental design given the number of attributes, attribute levels, choice questions, and design restrictions. D-optimality is used when orthogonality cannot be achieved in a design given the number of choice questions or the restrictions in the design. At a D-optimal point, the design has maximized the determinant of the information from the choice questions, and minimized the expected variance for the expected preference estimates.

Because Sawtooth Software does not allow for automated randomization of the sequence of questions within blocks, we took steps to avoid order effects that could potentially occur when individuals complete the same sets of questions in the same order. For each of the 6 blocks, we randomly ordered each set of 8 questions. When the survey is administered, each participant will be assigned to complete one of 48 sets of 8 questions so that all blocks are approximately equally represented.

To extract more choice information from individual respondents, we also will ask survey respondents whether they are willing to complete 4 additional choice questions for a total of 12 questions. To construct choice sets for the additional 4 choice questions, we generated a 24-question experimental design, again with no restrictions, and divided the design into 6 blocks of 4 questions. Each of these 6 blocks were replicated 4 times each with different question sequences. Respondents who are willing to complete 4 additional choice questions will be assigned to one of 24 sets of 4 questions. In total, the design includes 8 unique sequences for each of the 6 blocks in the 48-question design and 4 unique sequences for each of the 6 blocks in the additional 24-question design. To implement the scope test, we duplicated and recoded the risk levels such that either 10% or 20% would be shown as the highest risk levels. Figure 2 provides an overview of the experimental designs for the study.

Figure 2. Experimental Design



**not representative of actual sequences or block assignment*

1.1.4. Validity Tests and Quality Control

Specific modifications to the D-optimal design include validity tests to characterize the overall quality of the preference data collected in the study. Validity tests provide the opportunity to more closely evaluate responses that do not meet a minimum level of consistency with the economic principles underlying DCEs. However, the results must be interpreted carefully. Apparent inconsistencies could be accurate indicators of patient preferences or results of reasonable assumptions not accounted for in the study design (Janssen et al., 2018).

The study experimental design is expected to allow between-question consistency tests. These tests evaluate whether respondents who make choices that imply specific preferences also make choices that are consistent or inconsistent with such preferences elsewhere in the survey. Since these inconsistencies are correlated with measurement error, they can signal problems with the magnitude of the errors in the final preference estimates.

An additional indicator of internal validity is respondents' performance on comprehension questions included in the survey instrument after describing each study attribute, after the risk-grid tutorial, and after the practice-choice questions. These questions will be used to flag any respondents whose responses indicate that they may not completely understand the study attributes or question layout or may be generally inattentive to the survey content.

1.2. Ranking Question

To gain insights about device features that are inextricably linked to a specific type of pacemaker and thereby not amenable to inclusion in a DCE, a ranking exercise was designed to elicit information on participants' relative concern for these device features. These features include the insertion procedure, where the device is implanted/resides in the body, the need for leads, whether the device is externally visible, and duration of activity restrictions after the device is implanted (Table 2).

In a ranking exercise, it is important that all features included in the exercise are either positive or negative. Because the options pertaining to the insertion procedure, presence of leads and location are not clearly positive or negative, survey respondents are first asked to indicate which specific aspect of a pacemaker attribute is of greater concern for these three features (see first three rows of Table 2 below). For example, with regard to leads, respondents could choose "having a pacemaker without leads" or "having a pacemaker with leads", as shown in Table 2. The chosen feature of greater concern is then transferred to the ranking exercise so that all features shown in the ranking exercise represent negative aspects of pacemakers from the patient perspective.

Table 2. Items in the Ranking Exercise

Insertion procedure • Pacemaker inserted through cut in skin on the chest • Pacemaker inserted using tube through the groin
Leads • Having a pacemaker with leads • Having a pacemaker without leads
Pacemaker location • Having a pacemaker placed under the skin on the chest • Having a pacemaker attached to the walls inside the heart
No heavy lifting is allowed for 6 weeks (versus limited activity for 2 weeks)
Scar on chest (versus no scar)
Lump on chest (versus no lump)

2. STATISTICAL ANALYSIS

2.1. Data Summary

Descriptive statistics will be generated to describe the study sample. We will summarize the responses to each of the questions on respondents' demographic and socioeconomic characteristics and experience with heart problems and heart problem treatments, as well as questions testing whether respondents comprehended the information presented to them in the survey instrument. Summary statistics will be calculated based on the number of valid responses to the question and exclusive of any missing data. For categorical questions, we will compute the number and percentage of the sample providing each response. Continuous response questions will be summarized by the mean, standard deviation, minimum, and maximum values. [redacted] *P*-values will be computed using Pearson's chi-square test for categorical variables or Student's t-test for continuous variables.

2.2. Discrete-Choice experiment

Discrete-choice experiments generate complex cross-section/time-series choice data for each respondent. These data include a dichotomous dependent variable and are analyzed using advanced statistical methods. The basis for the analysis is the model specification assumed when generating the experimental design prior to survey implementation. That specification considers a categorical main-effects model for all study attributes. However, the specification assumed in the experimental design takes into account only statistical considerations. The statistical analysis of choices will provide a measure of the impact of changes in the attribute levels on the likelihood that treatments are selected by respondents, also referred to as attribute-level preference weights.

Respondents' reactions to the stimuli in DCE questions generally involve a complex decision process, and the final model specification must account for the pattern of choices as observed in the data set. For this reason, it is not advisable to pre-specify a final model specification before analyzing preference data. This statistical analysis plan lays out the initial analyses that will be conducted to assess the quality of the data and describes exploratory analysis strategies that will inform the final model specification.

The design and study specifications have been developed to support a robust analysis that includes estimates of relative preference weights and maximum-acceptable risk (MAR). In addition, the data will support more advanced analytic strategies to understand preference heterogeneity among respondents with different demographic, clinical, and treatment characteristics.

2.3. Evaluation of Data Quality and Its Impact on Treatment Preferences

The survey included 11 comprehension questions that can be used to identify any respondent who may not completely understand the study attributes or question layout or may be generally inattentive to the survey content. Also, the design of the DCE allows for several tests of internal validity (Johnson, Yang, & Reed, 2019). We describe below each test that will be used initially to evaluate data quality and then assess its effect on treatment preferences.

2.3.1. Respondents' Performance on Comprehension Questions

Respondents' performance on comprehension questions can be an indicator of internal validity. The questions will be used to identify any respondents whose responses indicate that they may not completely have understood the study attributes or question layout.

Along with the time spent completing the survey, we will use the information to determine whether some respondents should be excluded from the final analysis. In addition, we will assess whether respondents with more incorrect responses to comprehension questions exhibit statistically significantly different preference estimates compared to those who answered more comprehension questions correctly

2.3.1. Survey Completion

A respondent's data will be considered complete if that respondent has answered at least 1 discrete-choice question in the main DCE module. Respondents who do not meet this criterion cannot be included in the analysis because they provide no data on preferences.

2.3.2. Time to completion

The DCE survey is expected to take 20-30 minutes to complete; it includes several pages of reading materials and background questions in addition to the discrete-choice questions. We will assess the distribution of survey completion times and identify respondents that took too little time to complete the survey. We also will compare the time required to complete the survey between respondents who answered more or fewer comprehension questions correctly.

We will test the sensitivity of preference-estimate results to inclusion or exclusion of respondents who took less than 5 minutes to complete the survey since it is unlikely that an individual could read all the survey materials and give meaningful thought to their responses within that amount of time.

2.3.3. Response Non-Variance

The pacemaker alternatives in the choice questions are randomly assigned to first and second positions and there is no systematic relationship between the alternative placement and the attribute levels shown in each question. The probability that the preferred pacemaker alternative would appear in the same position for all 8 questions is less than 0.4%. For respondents answering 12 choice questions, the probability is just 0.02%. Therefore, we can infer that respondents who always select alternatives in the first or second positions are not attentive to the content of the choice questions. If response non-variance in the first or second position is observed for any respondent over the first 8 choice questions, their choice data will be excluded from the final analysis.

2.3.4. Attribute Dominance

In the DCE module, respondents choose among two pacemaker profiles in a series of choice questions. Some respondents may select the alternative with the better level of one attribute in all or most of these choice questions regardless of other attribute levels. When this response pattern is observed, it can indicate the respondent has a particularly strong preference for that attribute, and no combination of levels for the other attributes is sufficiently important to induce respondents to select an alternative with a less-preferred level of the dominant attribute. However, this pattern of dominated responses also could be evidence that a respondent simply made choices based on a single attribute to simplify the choice questions and thus avoid the effort of evaluating tradeoffs. It is not possible to definitively assess whether respondents with dominated response patterns expressed strong preferences for the dominant attribute or employed a simplifying heuristic; therefore, respondents who dominate on any attribute will be retained in the final sample.

We first will identify any respondents who selected the alternative with the better level of one attribute in all of the first 8 choice questions that have no overlapping levels, indicating dominated preferences for that attribute. We will report the numbers and percentages of respondents who dominated on each attribute. If more than 15% of the sample is observed to dominate on a single attribute, we will assess the impact of attribute dominance on the preference model estimates during the exploratory analysis phase (see Section 2.4). We also will test whether time spent completing the survey and responses in the background and treatment characteristics sections of the survey differed significantly between respondents who dominated and those who did not dominate. This analysis will be completed as part of the assessment of the background characteristics in latent-class analysis described in Section 2.4.4.

2.4. Specifying a Benefit-Risk Tradeoff Preference Model

Predetermining the specification of a regression model for analyzing choice data can lead to biased models, result in unnecessarily wide confidence intervals, and ultimately, result in missed opportunities to fully understand patients' preference patterns. To avoid these problems, this section outlines the initial analyses that will be conducted to evaluate relationships between respondents' choices and attribute levels, and it describes several potential exploratory analyses that will be conducted.

2.4.1. Initial Analyses

We will begin analyzing the choice data using a conditional logit model with effect-coded variables for all levels of each attribute. This model specification imposes the fewest assumptions regarding functional forms across preference weights within the study attribute levels but imposes simplifying assumptions on error terms.

Conditional-logit models run very quickly and often produce results similar to more complex models. We will use this initial model to identify any attributes and levels that require further investigation. For example, findings such as relationships between the estimated preference weights and levels within an attribute, interactions between attributes, or disordered preference weights (where clinically inferior levels are preferred over superior clinical levels) would require further exploratory analysis.

2.4.2. Scope Test

As explained in Section 1.1.2, we embedded a scope test to evaluate whether respondents are sensitive to the actual risk levels shown. Respondents assigned to the low-risk arm will see maximum risk levels of 10% for both the infection risk and complication risk attributes. Respondents assigned to the high-risk arm will see maximum risk levels of 20% for both risk attributes. If respondents systematically recode the risk levels qualitatively (for example, as low, medium, and high), the scope test would show that preference estimates are not sensitive to the overall risk range (i.e., the relative weight for the 10% level in the low-risk arm is not different than the relative weight for the 20% risk level in the high-risk arm). Because the other two levels in the experimental design overlap between arms, we can also assess whether the overall preference for the 1% and 5% risk levels are influenced by the maximum risk levels that respondents evaluate.

[redacted]

[redacted]

2.4.3. Subgroup Analyses

[redacted]

[2] Centers for Disease Control and Prevention. Defining adult overweight and obesity. https://www.cdc.gov/obesity/adult/defining.html. Accessed August 27, 2021.

- [redacted]

[redacted]

2.4.4. Final Preference Model

The final model specification will be estimated using a mixed-logit model with correlated random parameters (Train, 1998) where appropriate. We plan to specify normally distributed random-effects parameters to model preference heterogeneity. This model will be used to generate equivalence values described in Section 2.5.

2.4.5. Exploratory Preference Model

[redacted]

2.5. Equivalence Values

Preference-weight estimates from the final random-parameters logit model will be used to compute equivalence values in terms of risk-equivalent values and time-equivalent values as a means to compare the relative importance of preferred device features for the full cohort.

We will compute the maximum-acceptable risk (MAR) of infection and the MAR of complications that respondents would accept in exchange for their preferred pacemaker type, no discomfort, a device with longer battery life, and a device with more time since regulatory approval. We will compute MARs using the following steps:

1. [redacted]

[redacted]

The same approach described above can be used to estimate time equivalents where preference utility gains (i.e. log odds preference weights) for preferred pacemaker features are offset using equivalent reductions in preference utility gains for time attributes (i.e. reductions in years of battery life and smaller/larger number of years since regulatory approval). For example, this approach can be used to determine the maximum reduction in battery life that the average respondent would accept to avoid pacemaker-associated discomfort.

2.6. Predicted Choice Probabilities

Preference-weight estimates from the final random-parameters logit model will be used to compute (or simulate) the probability that the average respondent would choose a pacemaker profile defined by study attribute levels, compared with a profile with different study attribute levels.

[redacted]

Table 3. Pacemaker Profiles for Predicted Choice Probabilities

[redacted]

2.7. Ranking Exercise

To evaluate respondents' level of concern about device type-specific features, we will first report the numbers and percentages indicating which type of insertion procedure, the presence or absence of leads, and the location of the pacemaker is of greater concern. We also will report the number and percentage of respondents indicating that all three features associated with a transvenous pacemaker (i.e. insertion through chest, pacemaker with leads and pacemaker

[redacted]

placed under the skin on the chest) were of greater concern. We also will report the number and percentage of respondents indicating that all three features associated with a leadless pacemaker (i.e. insertion through a catheter, pacemaker without leads and pacemaker placed inside the heart) were of greater concern.

We then will compute the mean rank order for each device feature evaluated. For device features pertaining to procedure type, presence of leads and pacemaker location, the rank assigned to the type selected will be included in the mean estimate. When the feature type was not selected in the preceding question, the rank will be treated as missing. Therefore, the mean ranks will be conditional on the feature having been selected (or not available for selection- limited activity for 2 weeks, no scar on chest, no lump on chest).

We also will report the percentage of times that each device feature (without regard to which procedure type, presence of leads and pacemaker location was chosen) was chosen as the most concerning feature. Then, conditional on the number one ranking, we will report the percentage of times each type of device feature was chosen (e.g. when procedure type was the most concerning feature, the percentage of the time it represented insertion through the chest wall versus insertion through a catheter in the groin).

3. REPORTING

DCRI will conduct and report the results from the analyses described in this plan in a study report. DCRI will provide a presentation to Abbott to present a summary of the findings. DCRI also will develop a scientific manuscript for submission to a peer-reviewed journal to report on the study's findings.

4. REFERENCES

Bridges, J. F. P., Hauber, A. B., Marshall, D., Lloyd, A., Prosser, L. A., Regier, D. A., . . . Mauskopf, J. (2011). Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in Health, 14*(4), 403-413. doi:http://dx.doi.org/10.1016/j.jval.2010.11.013

Hensher, D. A., Rose, J. M., & Greene, W. H. (2008). 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), 126-133. doi:https://doi.org/10.1016/j.jtrangeo.2007.07.001

Janssen, E. M., Hauber, A. B., & Bridges, J. F. P. (2018). Conducting a Discrete-Choice Experiment Study Following Recommendations for Good Research Practices: An Application for Eliciting Patient Preferences for Diabetes Treatments. *Value in Health, 21*(1), 59-68. doi:https://doi.org/10.1016/j.jval.2017.07.001

Johnson, F. R., Yang, J.-C., & Reed, S. D. (2019). The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments. *Value in Health, 22*(2), 157-160. doi:https://doi.org/10.1016/j.jval.2018.07.876

Krinsky, I., & Robb, A. (1986). On Approximating the Statistical Properties of Elasticities. *The Review of Economics and Statistics, 68*(4), 715-719. Retrieved from https://EconPapers.repec.org/RePEc:tpr:restat:v:68:y:1986:i:4:p:715-19

Kuhfeld, W. (2010). *Marketing Research Methods in SAS*.

Reed Johnson, F., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., . . . Bridges, J. F. P. (2013). Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value in Health, 16*(1), 3-13. doi:https://doi.org/10.1016/j.jval.2012.08.2223

Train, K. E. (1998). Recreation demand models with taste differences over people. *Land Economics, 74*, 230 - 239.