Medical Device Innovation Consortium (MDIC) Patient Centered Benefit-Risk Project (PCBR)

Appendix A:

Catalog of Methods for Assessing Patient Preferences for Benefits and Harms of Medical Technologies

Medical Device Innovation Consortium

April 22, 2015

MDIC
MEDICAL DEVICE INNOVATION CONSORTIUM
ALIGN | ACHIEVE | ACCELERATE

TABLE OF CONTENTS

ABBREVIATIONS

AHP	analytic hierarchy process
BRAT	Benefit-Risk Action Team
CDRH	Center for Devices and Radiological Health
FDA	Food and Drug Administration
IMI	Innovative Medicines Initiative
MDIC	Medical Device Innovation Consortium
MI	myocardial infarction
PCBR	Patient Centered Benefit-Risk
PRO	patient-reported outcome
PROTECT	Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium
RTI-HS	RTI Health Solutions
SG	standard gamble
TTO	time tradeoff

1 INTRODUCTION

1.1 Background to the Patient Centered Benefit-Risk Initiative and the Catalog of Methods

There is increasing interest among regulators, industry sponsors, and patient-advocacy groups in ensuring that decisions regarding the development, regulatory approval, and reimbursement of pharmaceutical and medical technologies account for the views and preferences of patients. The Center for Devices and Radiological Health (CDRH) of the United States Food and Drug Administration (FDA) issued guidance in 2012 outlining the factors CDRH considers when assessing the benefits and harms of certain medical technologies during a premarket review (1). This guidance specifically emphasizes the role of the patient perspective and patients' risk tolerances in evaluating the balance between efficacy and safety of medical technologies. However, this guidance provides no clear direction for industry or for health authorities regarding how to collect or use patient-preference data in benefit-risk assessments. The goal of the Medical Device Innovation Consortium's (MDIC's) Patient Centered Benefit-Risk (PCBR) project is to help advance the regulatory science supporting the assessment of patient preferences for the benefits and harms of medical technologies.

The MDIC PCBR Steering Committee prioritized both the development of a framework for how patient-preference information can be used in regulatory benefit-risk assessments of medical devices and the development of a catalog of available methods to support this framework. The "Framework" was developed by the Framework Working Group, a subgroup of the PCBR Steering Committee members, supplemented by involvement of additional FDA reviewers, who represent those who would potentially be using the framework in the future. The Framework Working Group focused on developing a guide to help CDRH staff and applicants think about what patient-preference information is, when it might be useful in the regulatory process, how such information might be collected, other potential uses of patient-preference information, and what additional research might be valuable to improve the use of patient preference information in the regulatory process.

A Catalog Working Group was formed from the PCBR Steering Committee members and outside experts in preference assessment methodologies to develop a "Catalog" of the Methods that are available to assess patient preferences. Given the technical nature of the development of the Catalog of Methods, the MDIC PCBR Steering Committee sought external expertise for the development of the patient-preference methods catalog. RTI Health Solutions (RTI-HS), a health research organization with experience in health-preference studies, was contracted to develop the Catalog. RTI-HS then contracted with academic experts with specific expertise in the methods to be examined in the Catalog;

these experts also joined the Catalog Working Group. Table 1 presents the members of the Catalog Working Group.

Table 1. MDIC PCBR Project Catalog Working Group Members

Principal Investigator
▪ Brett Hauber, PhD, Senior Economist, Health Preference Assessment, RTI Health Solutions
RTI Health Solutions Staff
▪ Juan Marcos González, PhD, Senior Economist ▪ Angelyn Fairchild, Associate Research Economist ▪ Margaret Mathes, Medical Editor ▪ Kimberly Moon, Project Manager
Academic Experts and Consultants
▪ Scott Braithwaite, MD, MS, FACP, Department of Population Health, NYU School of Medicine ▪ Ken Deal, PhD, McMaster University ▪ James Dolan, MD, University of Rochester ▪ Martin Ho, MSc, FDA, Center for Devices and Radiological Health, Office of Surveillance and Biometrics ▪ Telba Irony, PhD, FDA, Center for Devices and Radiological Health, Biostatistics ▪ Ross Jaffe, MD, Versant Ventures and National Venture Capital Association (NVCA) ▪ Bennett Levitan, MD, PhD, Janssen R&D LLC, Johnson & Johnson ▪ Bryan Luce, PhD, MBA, Patient Centered Outcomes Research Institute (PCORI) ▪ Bray Patrick-Lake, MFS, Clinical Trials Transformation Initiative

FDA = Food and Drug Administration; MDIC PCBR = Medical Device Innovation Consortium Patient Centered Benefit-Risk.

The Catalog Working Group reported regularly to the PCBR Steering Committee during the development of the Catalog and members of the PCBR Steering Committee provided valuable input to the development of the Catalog. Additional input to and review of the Catalog was elicited from CDRH premarket reviewers and representatives from MDIC member companies. Dr. John F.P. Bridges, PhD, Associate Professor at the Johns Hopkins Center for Excellence in Regulatory Science and Innovation (CERSI), reviewed a draft of the Catalog and provided valuable feedback and guidance to the Working Group. Editorial support on an earlier draft of the Catalog was provided by Gail Zona from RTI-HS.

Financial support from an FDA BAA contract (HHSF223201400011C) made the Catalog and this Framework Report possible and was primarily used to fund the work of RTI-HS and outside experts on the Catalog.

1.2 Objective of the Catalog

The objective of the Catalog is to identify and provide an overview of a range of available patient-preference methods. The Catalog is intended to be a resource for researchers, industry sponsors, and FDA staff to consult when considering which patient-preference methods could be used when such data would be helpful in supporting development, regulatory, and postmarketing decisions related to medical technologies. Although the Catalog was developed, at least in part, in response to recent CDRH guidance (1), it is intended to be general enough to be a useful resource for sponsors, FDA staff, and other researchers considering the use of patient-preference methods in benefit-risk assessments of pharmaceuticals, biologics, and other health care products and services.

The Catalog is an introduction to a number of potential patient-preference methods. However, it is not a systematic review of methods. The Catalog is neither the definitive guide to determining which patient-preference method should be used in each situation nor a primer on how to implement each method. Rather, the Catalog is intended to provide an overview of methods and to be a starting point for understanding approaches to patient-preference studies.

1.3 Organization of the Catalog

Section 2 of the Catalog begins with a working definition of patient-preference methods and introduces the methods included in the Catalog. This section includes a discussion of the types of information provided by patient-preference methods, the categorization of methods in the Catalog, and a discussion of methods that were explicitly excluded from the Catalog. Section 2 concludes with a description of each method included in the Catalog. Section 3 begins by introducing questions to consider when evaluating patient-preference methods. The questions are intended to be a guide for understanding and distinguishing among methods. General considerations related to the implementation of a patient-preference study, regardless of the method used, are also discussed. In Section 4, each method is reviewed using the questions presented in Section 3. To the extent that they were identified by the Working Group, examples of prior use of each method are provided in Section 5. Section 6 presents suggestions for future research.

2 BENEFIT-RISK PREFERENCE METHODS

2.1 Definition of Patient-Preference Methods

Before determining the set of methods to include in the Catalog, the Working Group developed a working definition of patient-preference methods. The Working Group determined that patient-preference methods should allow for characterizing preferences for desirable attributes and acceptability of undesirable attributes and account for the relative nature of preferences (i.e., allow for direct or indirect comparison across attributes and, thus, treatment alternatives). The resulting working definition is as follows:

> *Patient preference methods are methods for collecting and analyzing data that allow quantitative assessments of the relative desirability or acceptability to patients of attributes that differ among alternative medical treatment approaches.*

2.2 Methods Included in the Catalog

2.2.1 Basic Principles for Including Methods in the Catalog

Before identifying methods to be included in the Catalog, the Working Group identified the following set of principles that could be used to guide the selection of methods for inclusion in the Catalog:

- The method should provide information on the relative importance of or tradeoffs among attributes that differ among alternative health interventions or diagnostic strategies, either directly or indirectly.
- The methodology, analysis, and interpretation of results of the method should have been published in peer-reviewed literature.
- The method should have been applied to health interventions previously.
- The method should be able to be applied to eliciting patient preferences even if the method is typically applied to elicit preferences or views of stakeholders other than patients.

These principles were developed to guide the process of selecting methods and were not treated as strict inclusion criteria. The final list of methods included in the Catalog was determined by consensus of the Working Group.

2.2.2 Quantitative Versus Qualitative Methods

Both qualitative and quantitative methods can be used to elicit information about patients' preferences for benefits and risks associated with medical technologies. Qualitative methods are designed to gain an understanding of patients' thoughts, feelings, and experiences. Although the concepts of interest are broadly defined before interacting with patients,

patients are encouraged to share and provide input without restrictions. The most commonly used qualitative methods are individual interviews and focus groups, although open-ended survey questions and social media also provide the opportunity to capture qualitative data. Typically, information gathered using qualitative methods is organized using some form of thematic analysis (see Fraenkel et al. [2] for an example). Although qualitative methods may also yield data that can be summarized numerically (e.g., the percentage of patients reporting a specific symptom, treatment benefit, or side effect), quantifying patient responses is not the primary objective of these methods (see Fried et al. [3] for an example of a qualitative study that yielded numeric data on patients' benefit-risk preferences). Quantitative data derived from qualitative studies may provide some information regarding the attributes of a medical technology that are important to patients and may even provide some evidence of the relative importance of these attributes to patients; however, quantifying the relative importance of attributes to patients using qualitative methods will most likely require supplementing a qualitative study with one or more of the quantitative methods presented in the Catalog.

Within the context of benefit-risk assessments, qualitative methods often are used to identify attributes that are important to patients in managing their disease and in evaluating treatment options. In fact, qualitative research, particularly concept elicitation, is often key to the development of rigorous studies designed to quantify benefit-risk preferences, and the development of many quantitative patient-preference studies relies heavily on qualitative research. Although qualitative methods may also provide an indication of patients' preferences among medical technologies, these methods are not optimal for quantifying the relative importance of individual treatment attributes or patients' willingness to trade off among these attributes.

In contrast to qualitative methods, quantitative methods are structured, with the type of data to be collected clearly defined and the response options limited to permit statistical analysis. For example, benefit-risk preference studies are explicitly designed to provide quantitative estimates of preference weights or the rate at which patients are willing to trade off among the benefits and risks of a medical technology.

Quantitative and qualitative methods need not be used in isolation and may actually prove most powerful when used in combination. For example, a survey that is primarily quantitative may include open-ended questions that provide supplemental information that can be analyzed using thematic analyses, and quantitative tasks, such as rating or ranking, may be included within a qualitative study to provide numerical outputs.

Although both qualitative and quantitative methods can be used to elicit patients' benefit-risk preferences, qualitative methods alone will likely not provide the level of information

required to inform regulatory benefit-risk assessments. Therefore, the Catalog focuses on quantitative methods for collecting and analyzing patient preference data.

2.2.3 Exclusion of Certain Potential Quantitative Methods From the Catalog

Several methods commonly used to incorporate the patient perspective in decision making are beyond the scope of the Catalog based on the guiding principles specified in Section 2.2.1. These include patient-reported outcome (PRO) methods and multicriteria decision-making methods. Strictly speaking, PRO methods are intended to measure health gains or losses that can only be assessed through direct reporting by a patient. In contrast, preference elicitation methods quantify how patients value those gains or losses. Moreover, quantitative benefit-risk trade-off preference methods are designed to quantify the value patients place on health outcomes or health care attributes whether or not the patients actually experience these outcomes. Multicriteria decision-making methods, such as multicriteria decision analysis (MCDA) and analytic hierarchy process (AHP), are intended to lead to or predict a decision. Eliciting preferences over the outcomes or health care attributes that define the decision problem is only part of these methods; the other part of these methods is combining preference weights with observed or expected outcomes in order to lead to or predict a decision. In other words, multicriteria decision-making methods require patient preference information as part of the inputs on how factors are weighed in these methods, but MCDA methods do more than elicit patient preferences. Therefore, neither PROs nor multicriteria decision-making methods are evaluated in the Catalog.

2.2.4 Descriptions of Methods Included in the Catalog

Table 2 presents the methods included in the Catalog. The methods are divided into four groups: structured weighting, health-state utilities, stated preference, and revealed preference. Although the grouping of methods may appear to be somewhat arbitrary, it reflects the nature of the method (e.g., qualitative vs. quantitative or stated vs. revealed), the current application of the method (e.g., whether used as part of a decision-analysis method or for the elicitation of preferences independent of the application of the results), and the underlying theoretic framework (e.g., decision-theoretic methods for most structured-weighting methods, expected-utility-theoretic methods for health-state-utility methods, and ordinal- or random-utility-theoretic methods for most stated-preference methods). This grouping scheme is meant only to facilitate a discussion of the methods and is not intended to preclude other grouping schemes that can be adequate in other contexts. In addition, some methods (e.g., simple direct weighting, ranking exercises, and point allocation) could be appropriately assigned to multiple groups.

Table 2. List of Methods Included in the Catalog

Group	Method
Structured-weighting	▪ Simple direct weighting ▪ Ranking exercises ▪ Swing weighting ▪ Point allocation ▪ Analytic hierarchy process ▪ Outranking methods
Health-state utility	▪ Time tradeoff ▪ Standard gamble
Stated-preference	▪ Direct-assessment questions ▪ Threshold technique ▪ Conjoint analysis and discrete-choice experiments ▪ Best-worst scaling exercises
Revealed-preference	▪ Patient-preference trials ▪ Direct questions in clinical trials

Table 3 describes each method and group of methods. Following each description is one or more citations of empirical examples of the use of the methods. Further detail on some of these examples is provided in Section 5 (Examples of Prior Use of Methods). If no empirical example is cited for a particular method, this is because a PubMed search in February 2015 did not reveal any published applications of this method to elicit patient preferences related to benefits and risks of pharmaceuticals or medical technologies.

Patient-preference methods can be used to provide one or more general categories of information: attributes, relative importance, or tradeoffs.

- Attributes
  - Attribute information indicates what matters to patients; that is, which attributes of a medical technology are important to patients when they weigh benefits and risks. Attributes are features that describe outcomes or events associated with treatment options (e.g., myocardial infarction [MI], nausea and vomiting, or response) or treatment characteristics (e.g., open surgery or recommended diet restrictions). They can be clinical in nature (e.g., outcomes or events) or nonclinical in nature (e.g., mode or frequency of administration or location of intervention or diagnostic service). Attributes can take on different levels representing the degree to which a treatment manifests, satisfies, or causes the attribute (e.g., probability or severity of the outcome or event). Attributes or attribute levels also can be combined into profiles.

  - Attributes can also be classified as health states (e.g., mild diarrhea, moderate pain, lung function), time in a health state (e.g., time to disease progression), or probability of a health state or the rate at which a health state occurs (patient rate, risk of serious adverse event). Attributes can be defined as the existence of the health state (e.g., severe diarrhea), a specific time in a health state (e.g., 1 week of severe diarrhea), or a specific probability or rate of a health state (e.g., 8% chance of severe diarrhea). Alternatively, attributes can be defined as a range of levels or change in the levels of a health state (e.g., severe diarrhea to mild diarrhea; 1 week of severe diarrhea to 2 days of severe diarrhea; 8% chance of severe diarrhea to 3% chance of severe diarrhea).
  - Throughout the Catalog, we use the term *attribute* to refer to a feature, characteristic, or health state. We use the term *levels* to refer to the various values an attribute can take and the term *profile* to describe a combination of attributes or attribute levels used to characterize a health state or medical technology.

- Relative importance
  - Relative importance information tells you how much each attribute matters to patients when compared with other attributes. Estimates of relative importance can be described as preference weights. Health-state utility studies typically provide weights for a health state without regard to the time in the health state or the probability of the occurrence of the health state. Structured-weighting methods are used in multicriteria decision methods, such as MCDA or AHP, and typically provide weights for a range of variations in health state that is determined to be relevant to the underlying benefit-risk decision. Stated-preference methods typically provide weights for a unit change in the range of an attribute or for the range of levels of the attribute included in the study design. All of these types of weights reflect the relative importance that patients place on different outcomes or features of a treatment; however, weights from these different methods can only be compared directly if the framing of the attribute (i.e., health state, time in a health state, probability of the health state) and the unit of measure (i.e., health state, change over a range of levels of the health state, or unit change in the health state) are comparable. Often relative importance weights estimated using different methods can be adjusted for framing and unit of measure to facilitate a direct comparison.

- Tradeoffs
  - Trade-off information tells you how much each attribute matters by explicitly stating what tradeoffs patients are willing to make to obtain or avoid a given attribute, or to change the level of a given attribute. Tradeoffs can be estimated between any pair of attributes or any pair of changes in attribute levels. The most common rates of tradeoff used in benefit-risk analysis are tradeoffs between health states (see Yuan et al. [4]) in the form of maximum acceptable risk (see Wilson et al. [5]) and minimum required benefit (see Ho et al. [6]). Although tradeoffs can be elicited indirectly and approximated by comparing the relative importance that patients assign to each attribute, obtaining accurate trade-off information may require quantitative methods designed explicitly for this purpose.

Data acquisition can be thought of as falling into three different categories: panel approaches, survey approaches, and clinical-study approaches. Panel approaches typically involve a small number of patients (perhaps 5-10) working together following a discussion guide or protocol and moderated by a facilitator to arrive at a mutual decision or consensus. Survey approaches typically involve relatively larger numbers patients (perhaps hundreds or thousands), and each patient responds independently to a structured set of questions. Clinical-study approaches involve decisions that result in or follow patients' exposure to a medical technology and observing effects or outcomes resulting from the exposure. In the Catalog, clinical-study approaches require a clinical study for implementation and do not include hybrid approaches in which health-state-utility or stated-preference surveys are included in a clinical study. In general, panel approaches are used for structured weighting, survey approaches are used for health-state utility and stated-preference methods, and clinical-study approaches are used for revealed-preference methods. However, there are exceptions to this general rule. For example, survey approaches can be used for many structured-weighting methods (see Hummel et al. [7] for examples and Stafinski et al. [8]). In addition, survey methods can be added to clinical studies to elicit preferences from a clinical-study sample.

Table 3. Descriptions of Methods Included in the Catalog

Group	Method	Brief Description
Structured weighting Structured weighting is the term used to describe the methods that typically are used to derive weights in multicriteria decision methods such as multicriteria decision analysis and analytic hierarchy process. Multicriteria decision methods typically are used to help people make evidenced-based decisions by systematically	Simple direct weighting	Simple direct weighting is a method for eliciting a weight for an attribute or attribute level on a predefined numeric scale. The scale is anchored between two defined points (e.g., death and perfect health, extremely important and not at all important, most and least). Higher ratings indicate greater weight. Each rating provides a single weight for an attribute or attribute level. Weights for different attributes or attribute levels can be compared directly as long as the different attributes or attribute levels are measured on the same numeric scale with the same anchors. (For an empirical example, see Stafinski et al. [8].) In special cases, when the anchors are defined such that 0 equals death and 1 equals perfect health, the resulting rating can be interpreted as the health state utility of an attribute or attribute level (see description below and Section 4.2). Although included among structured weighting methods in the Catalog, simple direct weighting could also be classified as a stated-preference method.
	Ranking	Ranking is a method for placing a set of attributes, attribute levels, or profiles in order of increasing or decreasing preference or importance. Ranking may be strict (no ties permitted) or nonstrict (ties permitted). Often, no numeric values reflecting the magnitude of preference are provided; however, methods exist for translating rankings into weights so that a relative weight can be applied to each feature, attribute level, or profile in the set (9-11). (For empirical examples, see Stafinski et al. [8] and Caster et al. [12].) Although included among structured weighting methods in the Catalog, ranking could also be classified as a stated-preference method.

Group	Method	Brief Description
combining clinical evidence with subjective judgments or weights. Structured-weighting methods in the Catalog are limited to those methods used in multicriteria decision methods to derive weights.	Swing weighting	Swing weighting requires that each attribute in a set be assigned a range of minimum to maximum levels, such that the full range of expected levels is included within the range. The attributes are then ranked in decreasing order of the importance that a change in each attribute, from its lowest level to its highest level, would have on a given decision. The attribute with the highest rank is assigned a weight of 100. The second attribute is then assigned a weight on a scale from 1 to 100, reflecting the degree a swing from its lowest to highest level would influence the decision, compared with the highest-ranked feature. Thus, higher weights indicate greater importance. This process is repeated for all attributes. The resulting weights are normalized to sum 100 and provide a weight for each attribute over the range of levels assigned to that feature.
	Point allocation	Point allocation requires that each attribute in a set be assigned points proportional to the importance associated with each attribute or specified changes in the levels of each attribute. The total number of points to be allocated among the attributes is fixed. The resulting values are usually normalized and interpreted as weights for the set of attributes. Higher weights indicate greater importance. (For an empirical example, see Stafinski et al. [8].) Although included among structured weighting methods in the Catalog, point allocation could also be classified as a stated-preference method.
	Analytic hierarchy process	Analytic hierarchy process requires that changes in the levels of each attribute are compared to derive weights that indicate the relative importance of changes in attribute levels to achieving a decision goal. This is accomplished through a series of pairwise comparisons between every pair of attributes. The questions are used to indicate which attribute is preferred, more likely, or more important as well as the strength of preference. Software interrogates a patient when choices are contradictory. These comparisons then are used to compute a weight for each attribute. For beneficial attributes, higher weights indicate greater preference. For undesirable attributes, higher weights indicate lower preference (13). (For empirical examples, see Hummel et al. [7] and Singh et al. [14].)

Group	Method	Brief Description
	Outranking methods	Outranking methods compare a set of decision alternatives or profiles using pairwise comparisons. Unlike swing weighting and AHP, outranking methods base their comparisons on whether one profile is at least as good as or better than the other. The individual comparisons are then aggregated to develop a ranking of profiles in terms of preference. The ranking is ordinal and does not provide a set of weights measured on a common scale. However, outranking methods are commonly combined with direct weighting in which outranking is used to compare the alternatives and the direct weighting is used to elicit weights.
Health-state utility Health-state utility methods yield an estimate of preferences for a health state (described as a single attribute or a profile) when compared with death and perfect health in which death is given a value of 0 and perfect health is given a value of 1. Higher weights equal greater preferences for a given health state.	Time trade-off	Time trade-off is a stated-preference approach in which patients are asked to choose between living a specified time in a specified state of health and a shorter time in perfect health. Health states can be defined by individual attributes or by profiles. The time spent in perfect health then is varied until the patient is indifferent between longer life in the worse health state and the shorter life in perfect health. The ratio of the shorter amount of time in perfect health to the longer amount of time in the health state is the health-state utility. Higher values indicate greater preference for the health state. (For an empirical example, see Avila et al. [15].)
	Standard gamble	Standard gamble is a stated-preference approach in which patients are asked to choose between a certain outcome and a gamble between two uncertain outcomes, each with a probability of occurrence and where their aggregate chance of occurrence is 100%. Typically, the certain outcome is described as a health state. Health states can be defined by individual attributes or by profiles. The two outcomes in the gamble are death and perfect health. The probabilities of death and perfect health are varied until the patient is indifferent between the certain outcome and the gamble between death and perfect health. The probability of perfect health at which the patient is indifferent between the certain outcome and the gamble is the health-state utility. (For empirical examples, see Avila et al. [15] and Kuchuk et al. [16].) Standard gamble also can be used to elicit risk tolerance directly. One minus the health utility can be interpreted as the maximum risk of death that would be tolerated in exchange for an improvement from experiencing the outcome to perfect health (i.e., elimination of the outcome or outcomes that define the health state). (For an empirical example, see O'Brien et al. [17].)

Group	Method	Brief Description
Stated preference Stated-preference methods are used to elicit relative preferences across attributes and changes in attribute levels using profiles. These methods differ from health-state utility methods because the weights elicited in these methods are not anchored on a cardinal scale between 0 and 1 in which 0 and 1 have a defined meaning. Instead, the weights elicited in stated-preference surveys can only be interpreted as ordinal rankings or	Direct-assessment questions	Direct-assessment questions ask patients to provide a direct answer to a statement or relative-importance question. The statement or question asks the patient the extent to which he or she agrees with the statement regarding an attribute or profile or whether he or she prefers or would choose one attribute or profile over all other available attributes or profiles. When a statement is structured to evaluate the extent to which a patient values an attribute or the importance a patient places on an attribute, the result is a weight for that feature. Most direct assessments of profiles provide only a preferred profile or an ordinal ranking of profiles from the set of profiles presented to a patient and, thus, do not result in weights. (For empirical examples of direct assessments of profiles, see Sarkissian et al. [18] and Yachimski et al. [19].)
	Threshold technique	The threshold technique is a stated-preference approach in which patients are asked to choose between a reference profile and an alternative profile. Both the reference profile and the alternative profile are defined by a common set of attributes, although the levels of each attribute can vary between the two alternatives. In the threshold technique, one attribute is considered to be the study object. If the reference profile is chosen, the level of the study object in the alternative profile is improved until the patient changes his or her choice from the reference profile to the alternative profile. If the alternative profile is chosen, the study object in the alternative profile is made worse until the patient changes his or her choice from the alternative profile to the reference profile. The point at which the patient switches his or her choice is the threshold. If the attribute that is the study object is a risk, the threshold probability is an estimate of maximum acceptable risk. If the attribute that is the study object is an efficacy or effectiveness outcome (i.e., benefit), then the threshold probability is an estimate of minimum required benefit. (For a list of empirical examples, see Hauber et al. [20]. For additional empirical examples, see Stafinski et al. [8], Kennedy et al. [21], and Kok et al. [22].)

Group	Method	Brief Description
relative preference weights.	Conjoint analysis and discrete-choice experiments	In discrete-choice experiments and conjoint analysis, the attributes of each medical treatment are assigned different levels that can be combined into profiles, and the profiles are combined into groups of profiles known as *choice sets*. The profiles and choice sets are determined by an experimental design. Each patient is presented with a series of choice sets and asked to choose one profile in each choice set. Alternatively, a patient could be asked to rank profiles in a choice set or rate his or her strength of preference for one profile over an alternative profile or to allocate the percentage of patients that would be treated best with each alternative profile in each choice set. The pattern of responses is analyzed to estimate the rate at which patients are willing to trade off among the attributes and changes in attribute levels included in the study. The results can provide measures of the relative importance of attributes or changes in attribute levels and the rate of trade-off among attributes or attribute levels. (For a list of empirical examples, see Hauber et al. [20]. For additional empirical examples, see Ho et al. [6], Mühlbacher and Bethge [23], Fraenkel et al. [24], Wouters et al. [25], and Guimaraes et al. [26].)
	Best-worst scaling	There are three types of best-worst scaling: object case, single-profile case, and multiple-profile case. In all cases, patients are presented with a set of alternatives and asked to identify the best or most important alternative and the worst or least important alternative. In the object case, attributes are combined into sets. Each set does not necessarily (and often does not) include all attributes. For each of a series of sets, patients are asked to indicate which of the attributes in the set is best or most desirable and which is worst or least desirable. In the single-profile case, each attribute takes on different levels. The attribute levels are combined into profiles. Patients are presented with a series of profiles and asked to indicate which attribute level in the profile is best or most desirable and which attribute level in the profile is worst or least desirable. In the multiple-profile case, attribute levels are combined into profiles, and the profiles are combined into sets of three or more. The multiple-profile case is very similar to a discrete-choice experiment. In each of a series of sets, patients are asked to indicate which profile is best or most desirable and which profile is worst or least desirable. In all three types of best-worst scaling, the pattern of responses is analyzed to estimate the relative importance of each attribute or attribute level. (For empirical examples, see Yuan et al. [4] and Peay et al. [27].)

Group	Method	Brief Description
Revealed preference Revealed-preference methods are used to analyze patients' choices and behaviors in the real world. These methods can provide information on the number of patients for whom the benefits of a medical technology outweigh the risks and potentially the reasons why patients believe that benefits outweigh risks; however, unlike stated-preference methods, revealed-preference methods often cannot be used to derive weights for or the relative importance of individual attributes or changes in attribute levels.	Patient-preference trials	Patient-preference trials are clinical trials in which patients are placed into arms of the trial depending on whether the patient has a strong preference for at least one of the medical technologies being studied. For example, in a trial with two medical technologies, A and B, patients with a strong preference for technology A are assigned to technology A, those with a strong preference for B are assigned to B, and those with no strong preference are randomly assigned to technologies A or B, effectively creating four study arms. An alternative approach is to randomize patients into two study arms. In the first arm, patients are randomly assigned to a medical technology. In the second arm, patients are assigned to a medical technology based on the patient's preference. If there are two technologies included in the study, then the study effectively has four arms. Follow-up questions can be administered to elicit the relative importance that each attribute of the medical technologies had on the patient's choice (or lack thereof). (For empirical examples, see Crowther et al. [28] and Long et al. [29].)
	Direct questions in clinical trials	Direct questions consist of questions that ask patients in a clinical trial to indicate their choice between a medical technology with which they have had experience and either their current care or an alternative technology. Direct questions can be administered in any phase of clinical research. The most rigorous approach to administering direct questions is to conduct a phase 3 trial with a classic cross-over design in which each patient is exposed to first one medical technology and then another; the patient then is asked to indicate which technology he or she would like to or plans to continue at the conclusion of the study. Follow-up questions can then be administered to elicit the relative importance a patient's experience with each attribute of the medical technologies had on his or her choice. (For empirical examples, see Escudier et al. [30] and Mitchell and Parikh [31].)

3 REVIEW OF METHODS

3.1 Questions to Consider when Evaluating Methods

The Catalog Working Group developed a set of questions to consider when evaluating patient-preference methods. The questions are divided into four categories: methodology-related questions, sample- or patient-related questions, analysis-related questions, and output-related questions. Table 4 presents the questions with descriptions. The questions are intended to be a guide for understanding and distinguishing among methods. For many methods, the questions can be answered in only general terms because there are multiple variations of a method and each variation might lead to a different answer to the question or because the answer to the question depends on the design or implementation of an individual study. Therefore, these questions are meant as a starting point for understanding patient-preference methods and their potential application to benefit-risk assessments. They are not intended to represent criteria for choosing a method for a specific application or an algorithm for conducting a patient-preference study.

Table 4. Questions to Consider

Questions	Description
Methodology-related questions	Methodology-related questions are intended to distinguish among the different patient-preference methods by considering some of the fundamental characteristics of the method such as data acquisition, hypothetical versus real-world decision making, the method for determining the attributes to be included in the study, and the experimental nature of the study.
How are the data acquired?	Data acquisition can be thought of as falling into 3 different categories: panel approaches, survey approaches, and clinical-study approaches. Panel approaches typically involve a small number of patients (perhaps 5-10) working together in a structured format to arrive at a mutual decision or consensus. Survey approaches typically involve relatively larger numbers patients (perhaps hundreds or thousands), and each patient responds independently to a structured set of questions. Clinical-study approaches involve exposing patients to a medical technology and observing effects or outcomes resulting from the exposure.
Are hypothetical scenarios required?	Hypothetical scenarios are often required to elicit patient preferences because real-world data do not provide sufficient information for evaluating tradeoffs among risks and harms of a medical technology. Some panel approaches and most survey approaches use hypothetical scenarios to elicit patient preferences. In general, clinical study approaches do not use hypothetical scenarios to elicit patient preferences.

Questions	Description
How are attributes and attribute levels determined and defined?	Attributes and attribute levels to be assessed in a patient-preference study can be determined prior to implementing the method (external identification) or as part of the method itself (internal identification). In most cases, panel methods typically use internal identification. Most survey approaches and most clinical study approaches typically use external identification to determine which attributes will be evaluated.
Is the method experimental?	Experimental methods are those patient-preference methods in which the researcher controls the attributes and scenarios to which patients are asked to respond. Typically, panel approaches are nonexperimental. Survey approaches can be experimental or nonexperimental. Clinical study approaches most often are experimental.
Sample-related questions	Sample-related questions are divided into 2 categories: sample size and patient burden. There is great variation in both the sample sizes that could be used for any method and the patient burden associated with any method. Therefore, these questions are intended to identify a range of possibilities for each method and to identify potential limitations of different methods with regard to each of these questions. Issues of sample frame, patient recruitment, response rates, or patient incentives are not addressed in the Catalog.[a]
What is the minimum sample size required?	Some methods require a sample of a minimum size to achieve the desired results of the study. Minimum sample size is meant to represent a lower bound on the sample size for any method. Typically, panel approaches do not require large sample sizes. Survey approaches often require a minimum sample size; however, there is a great deal of variation in the magnitude of the minimum sample size across survey methods. Sample sizes for clinical study approaches are often determined by power calculations that indicate the sample size necessary to test for the presence of a given effect.
What is the maximum sample size that can be reasonably achieved?	Panel approaches are typically implemented with smaller sample sizes (perhaps 5-10). Although these methods do not preclude the use of larger sample sizes, achieving larger samples may be difficult because of the nature of the methods. In contrast, the maximum sample size that is feasible with survey approaches or clinical study approaches may be limited only by considerations of time and cost.
What is the time commitment required of patients?	Different patient-preference methods require different levels of interaction with patients. Survey approaches may require the least amount of time by patients (probably measured in minutes). Panel approaches typically require relatively more patient time (probably measured in hours). Clinical study approaches may require the greatest time commitment if multiple site visits or interactions between the patient and investigator are required.

Questions	Description
What are the cognitive and knowledge requirements of patients?	Most patient-preference methods require patients to evaluate scenarios or make choices. The process of choosing among or evaluating alternatives often can be cognitively challenging. Different patient-preference methods may impose different levels of cognitive burden on patients and thus require a greater level of knowledge and/or cognitive capacity on the part of patients. For both ethical and cognitive reasons, patient-preference studies typically are not conducted among patients with diminished cognitive capacity or developmental insufficiencies or among vulnerable populations.
Analysis-related questions	Some patient-preference methods can be implemented and analyzed without the need for complex statistical analysis whereas other methods require advanced statistical techniques. The level of statistical sophistication required to analyze the data from any method often determines the need for specialized software, expertise to conduct the methods, and the ease with which the methods and results can be described.
Does the method require statistical analysis?	Some patient-preference methods yield results that are descriptive in nature or that can be presented as simple counts or proportions. Other methods require statistical inference and, therefore, require statistical methods. The complexity of statistical analysis required can also vary significantly across methods.
Does the method require specialized software?	Some patient-preference methods yield data that can be compiled and summarized without the need for specialized software. Some methods can yield data that can be analyzed using commonly available analytic software packages. Other patient-preference methods yield data that may require analysis using specialized software.
Can the results be described and interpreted easily?	Whether the results can be described or interpreted easily is typically a function of both the complexity of the analysis and the familiarity of end-users with the method. Some methods are simple and direct and can be easily interpreted. Some methods can yield results that are easily interpreted, but the process by which the results were achieved may be complex or lack transparency. Some methods may require statistical methods that may be difficult to explain and yield results that are not easily interpreted to those unfamiliar with the method.
Output criteria	Patient-preference methods can be used to provide 3 general types of information. ▪ Attributes: which attributes of a medical technology are important to patients when they weigh benefits and risks ▪ Relative importance: how much each attribute matters to patients ▪ Tradeoffs: how much each attribute matters and what tradeoffs patients are willing to make to obtain or avoid a given attribute.[b] In addition, patient preference methods can be used to quantify heterogeneity of preferences within a sample.[c]

Questions	Description
Can the method be used to identify attributes that are important to patients?	This type of information tells you what matters to patients; that is, which attributes of a medical technology are important to patients when they weigh benefits and risks.
Can the method be used to estimate weights for attributes?	This type of information tells you how much each attribute matters to patients. Obtaining this type of information requires using quantitative methods that provide a weight for each attribute.
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	This type of information tells you both how much each attribute matters and what tradeoffs patients are willing to make to obtain or avoid a given attribute. Understanding tradeoffs is usually necessary to calculate measures of maximum acceptable risk or minimum required benefit.[d]
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Preferences are heterogeneous. Some preference heterogeneity may be explained by differences in observable patient characteristics. Some preference heterogeneity may be unobserved either because the characteristics that explain heterogeneity are unobserved or because the characteristics that explain heterogeneity are latent or unknowable.

[a] See Section 3.2.2 for additional discussion of issues of representativeness and generalizability.

[b] See Section V of the Framework Report for additional discussion of these types of information and their uses in benefit-risk assessment.

[c] See Section 3.2.3 for additional discussion of preference heterogeneity.

[d] See Section II of the Framework Report and the glossary in Appendix B of the Framework Report for additional discussion of maximum acceptable risk and minimum required benefit.

3.2 General Considerations

When evaluating patient-preference methods, it is important to distinguish between the characteristics of the method and the implementation of the method. Many decisions regarding how a method is implemented can affect the results of a study. Sound methods can be implemented poorly and yield results of low quality or limited or no use. Therefore, any study of patient preferences should be conducted following good research practices to the extent that they exist. A detailed discussion of the proper implementation of each of these methods is beyond the scope of the Catalog; however, there are five overarching and interrelated issues that should be considered when evaluating the appropriateness of a method for use in assessing patient preferences: research question and study objective, representativeness of the sample and generalizability of results, the ability of the method to account for or describe within-sample heterogeneity of patients' preferences, validity of the study, and resources required to conduct the study.

3.2.1 Research Question and Study Objective

The first step in conducting a benefit-risk preference study is to define the research question based on the study objective. Research objectives can span the range from assessing patient views on a medical technology for guiding development decisions to developing statistical evidence of benefit-risk tradeoffs from a representative sample of patients to inform regulatory decision making. The study objective will vary across the medical product lifecycle. The research question should be designed to achieve the study objective and guide the choice of method. The research question and method will then determine the size of the sample required for any primary data collection. For example, obtaining patient views on specific benefits and potential harms of a medical technology early in development may require qualitative input or responses to simple surveys from a small group of patients or potential patients. Such studies typically can be conducted quickly and at relatively low cost. In this case, a representative sample probably is not necessary. In contrast, evidence of benefit-risk tradeoffs used to inform regulatory decision making may require a more robust statistical approach applied to a larger, more diverse, or more representative sample of patients to allow for the identification of subgroups with different preferences or to ensure generalizability of the results to a broader patient population.

The study objective and the specification of the research question will also determine the attributes that need to be evaluated in the benefit-risk study. The attributes of interest may be limited to those that have been identified as most concerning to patients or regulators or to those that differ among alternative medical technologies for the same indication. Alternatively, the attributes of interest may be determined by the primary endpoints and observed or expected safety signals in clinical trials. In each of these cases, the specific research question will be different and will depend on the study objective.

3.2.2 Representativeness of the Sample and Generalizability of Results

Representativeness of a sample is the extent to which the sample represents the population of interest on some selected characteristics of interest. The importance of the representativeness of the sample used in primary data collection will be determined by the study objective and should be specified in the research question. Generalizability of the results is the extent to which the results derived from a sample can be applied to the population of interest. Representativeness and generalizability are largely a function of sample size and the sampling frame. In general, it is more difficult to have representative samples and to generalize the results when sample sizes are smaller; however, larger sample sizes alone do not guarantee representativeness. In addition, results derived using a representative sample may be generalizable to a population on average; however, it may not be possible to analyze data from small subsamples of interest if the subgroup of interest represents a small proportion of the overall patient population. In addition, the population

from which the sample is drawn and the methods used to draw a sample from that population will influence the representativeness of the sample and the ability to generalize the results from the sample to the population of interest.

Representativeness is evaluated entirely by comparing the sample with the population of interest; thus, the representativeness of any sample will be determined, in part, by the research question. As in any area of research in which samples are used to gather data, the research question in a patient-preference study may involve understanding the preferences of a population with well-defined characteristics (e.g., a clinical trial population for which there are well-defined inclusion and exclusion criteria). In this case, recruiting a representative sample is relatively straightforward. However, the research question may be broader and involve understanding the preferences of the population that will be exposed to the medical technology in the future. In this case, recruiting a representative sample may be difficult because the characteristics of the overall population of interest may not be well understood. For example, a medical technology may be indicated to treat a given condition, but the number of patients with that condition and the distribution of ages and genders of patients in that population may not be known with any degree of certainty. Even if observable characteristics of the population are known, it is impossible to ensure that the preferences of any sample are representative of the overall population because differences in preferences may not be completely explainable by observable characteristics. Therefore, it is important to specify criteria for testing the representativeness of the sample that are relevant to the study and achievable.

Generalizability of the results to a population of interest almost always requires a representative sample. One exception arises when certain subgroups of patients are a small proportion of the overall population. In this case, it may not be possible to evaluate the preferences of the subgroup (or differences in preferences between the subgroup and the rest of the population) due to the small sample size of the subgroup. To remedy this, oversampling of this subgroup may be required. That oversampling will then lead to weighting of the sample for overall analyses and determining appropriate bases on which to calculate the weights.

Issues of representativeness and generalizability are not unique to preference studies. Patients in randomized controlled trials are rarely, if ever, representative of the patient population for which a medical technology will be indicated. Although inclusion criteria in clinical trials are intended to safeguard certain patients (e.g., pregnant women) or to reduce the likelihood of confounding (e.g., by excluding patients with comorbidities), these inclusion criteria have the effect of generating data for only a subset of the types of patients for whom the medical technology may be indicated. In addition, patient advice and testimony presented to advisory boards during the review process often is limited to a few

patients or a single patient group that may or may not be representative of the overall patient population for whom a medical technology may be indicated. Each of these situations is an example of a case in which representativeness may be lacking; however, in both of these cases, the information provided by samples that are not strictly representative of an overall patient population can be important to decision makers. Therefore, any study of patient preferences should include a transparent assessment of the representativeness of the sample in which the study was conducted.

3.2.3 Heterogeneity of Patients' Preferences

Patients' preferences for the benefits and harms of medical technologies are likely not consistent within populations. Sometimes preference heterogeneity can be attributed to differences in observable characteristics of subgroups of patients in the sample, such as age, weight, gender, and diagnostic variables. However, in most cases, the preference heterogeneity may simply reflect that different patients have different preferences and that those differences cannot be explained by differences in observable characteristics.

Most of the methods in the Catalog, with the possible exception of methods implemented using panel approaches where preference results are based on the consensus of a group, can provide some evidence of explainable preference heterogeneity to the extent that it exists as long as the sample includes a sufficient number of individuals with different characteristics to which preference heterogeneity can be attributed. Unexplained or latent preference heterogeneity can usually be identified even when there is little diversity of observable patient characteristics in the sample. For example, for methods that yield a single preference measure for each individual, such as the threshold technique that often results in a single risk threshold for each individual in the sample, the distribution of risk thresholds within the sample is a measure of preference heterogeneity across the sample whether or not that heterogeneity can be explained by observable characteristics. However, models that use multiple observations from each individual (e.g., conjoint analysis or best-worst scaling) do not always provide preferences at the individual level. In these cases, statistical methods can be used to measure preference heterogeneity. For example, one approach to analyzing data from a conjoint analysis is to use random-parameters logit, a statistical approach that provides a distribution around each preference weight estimated in the model (see Ho et al. [6] for an example). Latent-class analysis can be used to analyze conjoint-analysis data to quantify latent preference heterogeneity and identify segments of the sample composed of patients who have similar preferences that are distinct from the preferences of patients in other segments of the sample (32, 33).

Quantifying preference heterogeneity may be important if the benefits of a medical technology will outweigh the risks for some patients but not others, even if the expected clinical outcomes are the same among these patients. When all patients view the benefits of

a medical technology as exceeding the risks, understanding preference heterogeneity may be less important. However, if there are some patients who view the benefits of the medical technology as outweighing the risks and others who do not, knowing that such patients exist and being able to identify these two types of patients and the size of each of these groups relative to the overall sample will be necessary to infer the size of the overall patient population for whom the benefits might outweigh the risks. Even in cases in which the average patient in a sample perceives the risks to outweigh the benefits, it may be important to understand the size of the population for whom the benefits outweigh the risks (34) and, to the extent possible, identify specific characteristics that are correlated with the likelihood of being in one of these groups.

If preference heterogeneity exists, the subgroups of patients whose preferences are similar within the group and systematically different from other patients outside the group are referred to as *segments*. Whether membership of a segment is explained or unexplained, simulations can be used to identify the circumstances in which the benefits outweigh the risks for each segment. Optimal configurations of medical technologies may be substantially different among the different segments of the population, and adaptations of a medical technology could be designed for several segments. Such adaptations could provide greater net benefits to individual patients and to the overall patient population than would a single device developed for an average patient. Of course, the result of heterogeneity analyses may be that, although many segments can be served well with existing devices or adaptations of those technologies, some segments may have such distinctly different needs that evolutionary devices need to be developed specifically to provide benefits that outweigh the risks.

Prior experience with the medical technology under consideration or with the attributes included in a preference study (e.g., having a family member experience a stroke may increase the weight an individual associates with stroke) often may be the most likely patient characteristics that can explain preference heterogeneity across a sample. However, patient experiences can change over time, especially when patients are exposed to a medical technology or after years of experience with a disease, after which patients may develop adaptations that lessen the weight they associate with some aspects of the disease. As a result, measures of preferences for a single individual can change over time. Changes in preference measures over time typically can be explained by changes in experience over time (e.g., over the course of a clinical trial or before and after exposure to a medical technology). In theory, all the patient-preference methods included in the Catalog can be used longitudinally to assess changes in preferences over time.

Finally, preferences are influenced by culture and preferences may differ substantially among patients from different cultural groups. In this context, culture is meant to be a

broad term that encompasses the characteristics that a patient shares with members of a group who have a distinct, yet common, history, beliefs, or ways of life. Probably the most common characteristics that may indicate membership in a cultural group are country or area of residence, race or ethnicity, and national origin. To the extent that a reasonable number of distinct cultural subgroups can be identified within a sample, it may be possible to test for differences in preferences across cultures. However, it may not be possible to test for such preference heterogeneity when samples include relatively small samples of people from each of a large number of countries, as is not uncommon in clinical trials.

3.2.4 Validity of Patient-Preference Methods

Validity is the extent to which quantitative measures of relative importance or tradeoffs reflect the true preferences of patients. Validating patient-preference assessments is intrinsically difficult for stated-preference methods because these methods typically involve scenarios in which patients are asked to make hypothetical choices without actually experiencing the consequences of that choice. Therefore, it is unknown whether patients would actually do what they say they would do. Often hypothetical choices are necessary because observing actual choices is impossible or observing actual choices does not provide sufficient variation in attributes or attribute levels to tease out the rates at which patients would be willing to trade off among attributes.

Although there is no clear definition of what constitutes a valid patient-preference study, there are methods for evaluating patient-preference data. One method includes examining the consistency of responses that patients provide when asked a series of preference questions Consistency reviews can be based on responses to repeated questions or whether responses to a series of questions reflect properties of monotonicity (more of a good thing is better than less, and less of a bad thing is better than more) and transitivity (if a patient prefers A to B and B to C, then that patient should prefer A to C). However, it may be incorrect to assume that inconsistencies invalidate a study. First, no study will be free from inconsistencies; however, there is not a standard level of inconsistency against which the validity of a study can be assessed. Second, inconsistencies that may appear to represent errors, irrational responses, or responses that lack face validity may indeed be rational (35). Therefore, users of patient-preference data must use caution when using consistency measures such as these to evaluate the quality of a patient-preference study.

Two additional methods for assessing the validity of a patient-preference study or a patient-preference method may be to examine the test-retest reliability of the patient-preference method or patient-preference instrument. One study found positive test-retest reliability of time tradeoff (TTO) (36). However, there is little evidence regarding the test-retest reliability of a large majority of stated-preference survey instruments, as these instruments are often developed for a single study and not used again in subsequent studies. In

addition, test-retest reliability may be difficult to establish for preference methods, as patient preferences may change over time due to factors that are unexplained and thus cannot be controlled in such a study. Face validity is perhaps an easier test of validity. In some cases, errors in the ranking or weighting of widely disparate outcomes may be an obvious violation. For example, if a serious outcome such as disabling stroke is ranked as less important than a minor outcome such as the common cold, a study probably would be considered to lack face validity. In addition, the ranking of preference weights for naturally ordered outcomes may provide some information regarding face validity. For example, if achieving a lower response rate is rated as more important than achieving a higher response rate or if a more severe adverse event is rated as being better than a less severe adverse event, a study could be considered to lack face validity. However, as noted above, caution is needed in evaluating what appear to be violations of face validity except in certain extreme or obvious cases (35).

Assessing the ability of preference data to predict a patients' actual or hypothetical decision may be the most obvious method for evaluating the validity of a patient-preference study. When patients are asked a series of questions, the resulting preference estimates can be used to predict how patients would respond to any individual question in the series (37, 38). Although this information may provide insight into the ability of a patient-preference study to predict patient choices, these results should be viewed with caution. First, there is no standard by which to judge what level of predictive accuracy is necessary for a study to be considered valid. Second, many patient preference studies are designed to evaluate the tradeoffs patients are willing to make among attributes and prediction is not the objective of the study. Therefore, using the precision with which the data from a patient preference study can predict a single decision to evaluate the study may be inconsistent with the underlying objectives of the study.

It may be possible to gain insight into the validity of stated preferences by evaluating the same research question using two or more different preference methods. Although this type of analysis would not provide a formal test of validity, if the results derived using two or more methods are similar or support the same conclusion, the user of these results would have greater confidence in the validity or accuracy of the methods. To date, a few studies comparing methods have been published (39-41); however, these studies have been designed to understand the properties of the methods and have not necessarily been used to validate the results of one study using the results obtained using a different method.

Some may consider assessing the ability of the results of a patient-preference study to predict a patients' actual decision in the real world to be the ultimate test of validity. However, opportunities to make such a comparison are limited because not all patients have the opportunity to choose among alternatives in the real world because only one or a limited

number of alternatives exist or because the choice is made on behalf of the patient by a physician or a third-party payer. However, a patient-preference study designed to generate information that can help validate the evidence collected through premarket patient-preference studies and to understand the effect of experience on patient preferences could be conducted. One study could be conducted among patients who would be eligible for a medical technology but who have no prior experience with the medical technology. The same patient-preference study could then be conducted among patients who receive the medical technology once the medical technology becomes available. There may also be different study designs that could capture similar information. To date, we are unaware of any published study that has used this or a similar approach.

3.2.5 Resources Required to Conduct a Patient-Preference Study

Patient-preference methods differ in their approaches to acquiring data and in the complexity of the methods required to analyze the data. These differences result in variations in the time and cost required to complete a patient-preference study. It is impossible to generalize about the time of financial resources required to conduct a patient-preference study by method alone because so much of the time and the cost of conducting a study depends on the study implementation. Instead, it may be more informative to describe broadly the steps required to implement each approach to patient-preference measurement – panel approaches, survey approaches, and clinical-study approaches.

Panel approaches require developing the materials required to conduct the panel, recruiting and scheduling panel participants, conducting the panel, and analyzing the data. Although conducting the panel may take only a day, the preparation of the panel materials and the recruiting and scheduling of panel participants could take weeks or months. Also, although compiling and analyzing the results of the panel may be relatively straightforward, experience is required to develop panel materials and conduct the panel.

Survey approaches require developing the survey instrument, recruiting patients, administering the survey and analyzing the results. The time required to develop a survey instrument and analyze the data will depend upon the type of survey and the planned use of the results of the survey. Simple surveys such as those using simple direct weighting, ranking, rating, or direct assessment questions for which summary statistics such as counts, means, standard deviations and proportions are reported might be developed more quickly than more complex surveys using more sophisticated methods, which typically take at least several months. The use of the survey results may also dictate the level of rigor required and expertise for survey development. Developing surveys used to generate data for internal decision-making purposes may require less time and effort than developing a survey to generate evidence to support a regulatory submission.

Recruiting patients and administering a survey can have a significant impact on both time and cost of a patient-preference study. Recruiting patients through a patient organization or existing Internet panel may require less time and cost than recruiting patients through referral from healthcare providers. In addition, recruiting patients with rare diseases may require more effort than recruiting patients with highly prevalent diseases. Finally, Internet surveys may require less time and cost than face-to-face interviews or pencil-and-paper surveys.

Clinical-study approaches probably require the most time and cost. Interventional and observational clinical studies are subject to well-established good practices and regulations to which panel and survey methods may not be.

Further discussion of the factors related to the resources available to undertake a patient preference study is provided in Section V of the Framework Report.

3.3 Reviewing Different Patient-Preference Methods Relative to the CDRH Weight-Loss Preference Study

Ho et al. (6) described a patient-preference method designed to elicit preferences of obese people for weight-loss technologies. The study was a pilot study sponsored by CDRH and was motivated by the desire to assess patient preferences over a range of potential weight-loss devices, each with different levels of potential benefits and potential risks. The attributes and attribute levels included in the CDRH study were determined by consensus of clinical, regulatory, and methodology experts based on knowledge of existing and potential devices at the time the study was conducted. The study was designed to generate patient preference information that could assist CDRH in evaluating patients' risk tolerance for weight-loss devices. The relevant information included patient preferences on device efficacy and safety and device-related process measures. The study elicited preferences for several types of device benefits to determine the minimum clinically meaningful benefit required for patients to accept a given risk profile. The attributes included in the study were:

- Percentage of weight loss
- Duration of weight loss
- Reduction in the need for prescription drugs to treat obesity-related comorbidities or reduction in the risk of developing obesity-related comorbidities in the future
- Risk of dying as a result of getting the device
- Risk of device-related side effects requiring hospitalization
- Duration of side effects
- Dietary restrictions
- The type of operation required to implant the device

One finding from the CDRH weight-loss study was that, all else equal, a 30% reduction in total body weight was approximately 1.3 times as important as avoiding a 1% risk of death from the device. The results of the study also implied that, for a 30% reduction in total body weight, patients would, on average, be willing to accept a 1.4% risk of death due to the device.

A discrete-choice experiment was chosen as the method for eliciting patient preferences for weight-loss devices in the CDRH study for a number of reasons. First, discrete-choice experiments have been widely used to elicit patients' benefit-risk preferences and to estimate the maximum level of treatment-related risk that patients would be willing to accept to achieve a given benefit (MAR) and the minimum level of treatment benefit that

would be required for patients to accept a given level of treatment-related risk (20). Second, discrete-choice experiments can be used to estimate both tradeoffs and relative importance. Finally, discrete-choice experiments can be used to explore the distribution of patients' benefit-risk preferences across a sample and, therefore, provide information regarding the proportion of the population for whom the benefits of a technology are likely to exceed the risks.

Although a discrete-choice experiment was used in the CDRH weight-loss preference study, other methods could potentially have been used to address the same research question. Therefore, following the review of each patient-preference method in Section 4 of the Catalog, we provide a brief, top-level description of how each method might have been used to elicit relative importance of and tradeoffs among the potential benefits and harms of weight-loss devices. Examples are not provided for conjoint-analysis methods or best-worst scaling because these methods are similar to those used by Ho et al. (6). In addition, there are now a few examples comparing the use of discrete-choice experiments and best-worst scaling in patient-preference studies (see Hollin et al. [39]). Examples are not provided for TTO and standard gamble (SG) because using these methods requires a different conceptual framework for conducting benefit-risk analysis.

Each example is intended only to illustrate how a patient-preference method might be applied to eliciting preferences for the attributes included in the CDRH weight-loss study. The examples are not case studies and are not intended to provide a "how-to" guide to conducting studies of patients' preferences for the benefits and harms of weight-loss devices using these methods. In addition, a discussion of the advantages and limitations of each method in addressing the weight-loss research question is not evaluated systematically. Rather, the examples are designed to provide additional information and context to potential users of these methods with a starting point for thinking about how different methods can be applied to an existing patient-preference question. As is indicated in suggestions for future research in Section 6 of the Catalog, we recommend that additional studies be conducted in which multiple patient-preference methods are used to address the same research question. Such studies will enable users to directly compare and contrast the performance of different patient-preference methods and their implications for decision making, along with the relative advantages and limitations of any given method in a different situation.

4 METHODS REVIEWS

4.1 Structured-Weighting Methods

Structured weighting is the term used to describe the methods that typically are used to derive weights in multicriteria decision methods such as multicriteria decision analytic methods and AHP. Multicriteria decision methods typically are used to help people make evidenced-based decisions by systematically combining clinical evidence with subjective judgments or weights. Structured-weighting methods in the Catalog are limited to those methods used in multicriteria decision methods to derive weights (see Dolan [42], Felli et al. [43], and Dodgson et al. [44]). Structured-weighting methods included in the Catalog are simple direct weighting, ranking, swing weighting, point allocation, AHP, and outranking.

4.1.1 Simple Direct Weighting

Overview

Simple direct weighting is a method for eliciting a weight for an attribute or attribute level on a predefined numeric scale. The scale is anchored between two defined points (e.g., death and perfect health, extremely important and not at all important, most and least). Higher ratings indicate greater weight. Each rating provides a single weight for an attribute or attribute level. Weights for different attributes or attribute levels can be compared directly as long as the different attributes or attribute levels are measured on the same numeric scale with the same anchors.

In special cases, when the anchors are defined such that 0 equals death and 1 equals perfect health, the resulting rating can be interpreted as the health state utility of an attribute or attribute level.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are typically used; panel approaches can be used
Are hypothetical scenarios required?	Panel methods typically used to evaluate real-world scenarios; survey methods typically used to evaluate hypothetical scenarios
How are attributes determined and defined?	Survey approaches typically require external identification; panel approaches can allow for internal identification
Is the method experimental?	Experimental design not required
Sample criteria	
What is the minimum sample size required?	No minimum sample size is required unless statistical inference is required
What is the maximum sample size that can be reasonably achieved?	No maximum sample size for survey approaches; large sample sizes may be cost- and time-prohibitive for panel approaches
What is the time commitment required of patients?	Minutes to less than an hour required for survey approaches; time for getting to and attending the panel session required for panel approaches; preparation for panel session may also be required
What are the cognitive and knowledge requirements of patients?	Minimal cognitive requirements for survey approaches; panel approaches may be more challenging; understanding of attributes required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used in certain circumstances
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Panel approaches can allow for internal identification; survey approaches can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, simple weights for all attributes
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Possibly: ratios of weights could be interpreted as tradeoffs; however, tradeoffs are not elicited directly
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Possibly: Using survey approaches, differences in weights across individuals in a sample can be characterized; if panel approaches are used to achieve consensus among the panel, then this is not possible; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: Each weight-loss device attribute or attribute level could be evaluated independently by asking patients to rate each attribute or attribute level on a numeric scale. For example, a rating question could be presented as follows: "On a scale of 0 to 10, where 0 is totally unimportant and 10 is extremely important, how important is a 30% reduction in total body weight to you when you are thinking about getting a weight-loss device?" The question could then be repeated for each attribute or attribute level to provide a set of importance weights for all attributes or attribute levels. For example, suppose a 30% reduction in total body weight loss was assigned a value of 10. If the patients completing the direct-weighting exercise had preferences that were similar to those patients who participated in the CDRH study, we would expect that patients would assign a value of approximately 8.1 to a 1% risk of death due to the device. The weights can then be compared across attributes or attribute levels. The weights could also be used to determine a subset of the most important attributes of weight-loss devices when only a limited number of items can be considered. For example, a user of this information could determine the five most important attributes to consider in a separate preference elicitation effort or moving forward with device development decisions.

4.1.2 Ranking Exercises

Overview

Ranking is a method for placing a set of attributes, attribute levels, or profiles in order of increasing or decreasing preference or importance. Ranking may be strict (no ties permitted) or nonstrict (ties permitted). Often, no numeric values reflecting the magnitude of preference are provided; however, methods exist for translating rankings into weights so that a relative weight can be applied to each feature, attribute level, or profile in the set.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches and panel approaches can be used
Are hypothetical scenarios required?	Panel methods typically used to evaluate real-world scenarios; survey methods typically used to evaluate hypothetical scenarios
How are attributes determined and defined?	Survey approaches typically require external identification; panel approaches can allow for internal identification
Is the method experimental?	Experimental design not required
Sample criteria	
What is the minimum sample size required?	No minimum sample size if weighting not required
What is the maximum sample size that can be reasonably achieved?	No maximum sample size for survey approaches; large sample sizes typically cost- and time-prohibitive for panel approaches
What is the time commitment required of patients?	Minutes to less than an hour required for survey approaches; time for getting to and attending the panel session required for panel approaches; preparation for panel session may also be required
What are the cognitive and knowledge requirements of patients?	Minimal cognitive requirements for survey approaches; panel approaches may be more challenging; understanding of attributes required
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used in certain circumstances. Weighting requires more advanced statistical methods
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used

Criteria	Review
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret for descriptive statistics. Weighting methods are more difficult to describe; however, results are easy to describe and interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Panel approaches can allow for internal identification; survey approaches can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Possibly: simple weights for all attributes can be estimated only if weighting methods are applied to ranking
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Possibly: if weighting methods are applied, ratios of weights could be interpreted as tradeoffs; however, tradeoffs are not elicited directly
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Possibly: using survey approaches, differences in weights across individuals in a sample can be characterized; if panel approaches are used to achieve consensus among the panel, then this is not possible; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: A list of weight-loss device attributes or attribute levels could be presented to patients. Patients would then be asked to rank the attributes or attribute levels from best to worst or most important to least important. For example, if the patients completing the ranking exercise had preferences that were similar to those patients who participated in the CDRH study, we would expect that patients would, on average, rank a 5% risk of death due to the device as most important and changes in the risk of side effects requiring hospitalization as least important. The results of this type of question will not provide numeric weights for the attributes or attribute levels unless the ranking is translated into weights using mathematical methods. However, the proportion of patients assigning each rank to each attribute or attribute level could be reported.

4.1.3 Swing Weighting

Overview

Swing weighting requires that each attribute in a set be assigned a range of minimum to maximum levels, such that the full range of expected levels is included within the range. The attributes are then ranked in decreasing order of the importance that a change in each attribute, from its lowest level to its highest level, would have on a given decision. The attribute with the highest rank is assigned a weight of 100. The second attribute is then assigned a weight on a scale from 1 to 100, reflecting the degree a swing from its lowest to highest level would influence the decision, compared with the highest-ranked feature. Thus, higher weights indicate greater importance. This process is repeated for all attributes. The resulting weights are normalized to sum 100 and provide a weight for each attribute over the range of levels assigned to that feature.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Panel approaches are typically used; survey approaches can be used
Are hypothetical scenarios required?	Panel methods typically used to evaluate real-world scenarios; survey methods can be used to evaluate hypothetical scenarios
How are attributes determined and defined?	Panel approaches typically use internal identification; survey approaches may require external identification
Is the method experimental?	Experimental design not required
Sample criteria	
What is the minimum sample size required?	No minimum sample size
What is the maximum sample size that can be reasonably achieved?	No maximum sample size for survey approaches; large sample sizes typically cost- and time-prohibitive for panel approaches
What is the time commitment required of patients?	Minutes to less than an hour required for survey approaches; time for getting to and attending the panel session required for panel approaches; preparation for panel session may also be required
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements for survey approaches; panel approaches may be more challenging; understanding of attributes and tradeoff task required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Basic calculations are required; basic regression methods can be used in certain circumstances
Does the method require specialized software?	Basic spreadsheet software is required; commonly available statistical packages can be used
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Panel approaches can allow for internal identification; survey approaches can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, weights for all attributes
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Yes, pairwise tradeoffs estimated for all attributes
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Possibly: Using survey approaches, differences in weights and tradeoffs across individuals in a sample can be characterized; if panel approaches are used to achieve consensus among the panel, then this is not possible; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: A list of weight-loss device attributes or attribute levels could be presented to patients. Patients would then be asked to rank the attributes or attribute levels from best to worst or most important to least important when choosing a weight-loss device. Each attribute is assigned a relevant range of levels. Using panel approaches, the panel might determine the relevant range of levels for each attribute. Using survey approaches, the research team likely would need to determine the range of levels. Suppose that the relevant range of levels for total body weight loss is determined to be from 0% to 30% and the range of levels of the risk of dying as a result of getting the device is from 0% to 1%. Once the range of levels is determined, patients would be asked to provide a rating from 1 to 100 reflecting the relative importance of changing a weight-loss device from having the worst or least desirable level to the best or most desirable level in the range of each attribute individually. In this rating, 100 is given to the most important change in the levels of a single attribute. If the patients completing the rating exercise had preferences that were similar to those patients who participated in the CDRH study, we would expect that patients would rate a 30% reduction in total weight loss to be tied with a 60-month duration of weight loss. If both attribute changes are considered to be the most important when choosing a weight-loss device, a 30% reduction in total body weight and a 60-month duration of weight loss would be assigned a score of 100. Then the patient would be asked to assign a weight indicating the importance of each of the other attribute changes, including a change in the risk of death from 0% to 1%, relative to the importance of a 30% reduction in total body weight. We would expect that the score assigned to a 1% change in the risk of death in this case would be 81. The weights make possible the comparison of attributes and attribute levels. The weights could also be applied to the characteristics of alternative devices to provide a measure of the extent to which one combination of attributes or attribute levels would be preferred to an alternate combination of attributes or attribute levels.

4.1.4 Point Allocation

Overview

Point allocation requires that each attribute in a set be assigned points proportional to the importance associated with each attribute or specified changes in the levels of each attribute. The total number of points to be allocated among the attributes is fixed. The resulting values are usually normalized and interpreted as weights for the set of attributes. Higher weights indicate greater importance.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches and panel approaches can be used
Are hypothetical scenarios required?	Panel methods typically used to evaluate real-world scenarios; survey methods typically used to evaluate hypothetical scenarios
How are attributes determined and defined?	Panel approaches typically use internal identification; survey approaches typically require external identification
Is the method experimental?	Experimental design not required
Sample criteria	
What is the minimum sample size required?	No minimum sample size
What is the maximum sample size that can be reasonably achieved?	No maximum sample size for survey approaches; large sample sizes typically cost- and time-prohibitive for panel approaches
What is the time commitment required of patients?	Minutes to less than an hour required for survey approaches; time for getting to and attending the panel session required for panel approaches; preparation for panel session may also be required
What are the cognitive and knowledge requirements of patients?	Minimal cognitive requirements for survey approaches; panel approaches may be more challenging; understanding of attributes required
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used in certain circumstances
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret

Criteria	Review
Output criteria	
Can the method be used to identify attributes that are important to patients?	Panel approaches can allow for internal identification; survey approaches can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, simple weights for all attributes
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Possibly: Ratios of weights could be interpreted as tradeoffs; however, tradeoffs are not elicited directly
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Possibly: Using survey approaches, differences in weights across individuals in a sample can be characterized; if panel approaches are used to achieve consensus among the panel, then this is not possible; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: A list of weight-loss device attributes or attribute levels could be presented to patients. Patients would then be asked to allocate a fixed number of points (assume 100 points) across the set of attributes or attribute levels where more points imply greater importance of an attribute in the decision to choose a weight-loss device. Because patients are endowed with a fixed number of points, their point allocation must add up to the total number of points given across all attributes. If the range for the risk of death was 0 to 1%, and the preferences of these patients are similar to those who participated in the CDRH study, we would expect patients, on average, to allocate approximately 15 points to a 1% risk of death due to the device and approximately 18 points to a 30% reduction in weight loss. The weights can then be compared across attributes or attribute levels. The weights could also be applied to the characteristics of alternative devices to provide a measure of the extent to which one combination of attributes or attribute levels would be preferred to an alternate combination of attributes or attribute levels.

4.1.5 Analytic Hierarchy Process

Overview

Analytic hierarchy process requires that changes in the levels of each attribute are compared to derive weights that indicate the relative importance of changes in attribute levels to achieving a decision goal. This is accomplished through a series of pairwise comparisons between every pair of attributes. The questions are used to indicate which attribute is preferred, more likely, or more important as well as the strength of preference. Software interrogates a patient when choices are contradictory. These comparisons then are used to compute a weight for each attribute. For beneficial attributes, higher weights indicate greater preference. For undesirable attributes, higher weights indicate lower preference.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Panel approaches are typically used; survey approaches can be used
Are hypothetical scenarios required?	Panel methods typically used to evaluate real-world scenarios; survey methods can be used to evaluate hypothetical scenarios
How are attributes determined and defined?	Panel approaches typically use internal identification; survey approaches may require external identification
Is the method experimental?	Experimental design not required, but could be used
Sample criteria	
What is the minimum sample size required?	No minimum sample size
What is the maximum sample size that can be reasonably achieved?	No maximum sample size for survey approaches; large sample sizes typically cost- and time-prohibitive for panel approaches
What is the time commitment required of patients?	Minutes to less than an hour required for survey approaches; time for getting to and attending the panel session required for panel approaches; preparation for panel session may also be required
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements for survey approaches; panel approaches may be more challenging; understanding of attributes and tradeoff task required
Analysis criteria	
Does the method require statistical analysis?	Advanced statistical analysis often required

Criteria	Review
Does the method require specialized software?	Commonly available statistical packages can be used; specialized software available
Can the results be described and interpreted easily?	Methods and results can be difficult to describe; advanced statistical methods make describing results difficult; combining positive and negative values for weights can make results difficult to interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Panel approaches can allow for internal identification; survey approaches can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, weights for all attributes
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Yes, pairwise tradeoffs estimated for all attributes
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Possibly: Using survey approaches, differences in weights and tradeoffs across individuals in a sample can be characterized; if panel approaches are used to achieve consensus among the panel, then this is not possible; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: Each patient would be presented with a pair of attributes and asked to indicate the relative strength of preference for the attributes when choosing a weight-loss device using a visual analog scale. If we assume that the pair of attributes comprises a 30% reduction in total body weight and a 1% risk of death due to the device, the patient would be asked to rate these two attributes between one end of the scale indicating that a 30% reduction in total body weight is very important and a 1% risk of death is completely unimportant and the other end of the scale indicating that a 30% reduction in total body weight is completely unimportant and a 1% risk of death is very important. The visual analog scales are often numeric and symmetrically depict the intensity of relative preferences for attributes around zero. If the patient gives a rating of zero for a pair of attributes, then the patient is indicating that he or she is indifferent between the two attributes or that the attributes are of equal importance when choosing among weight-loss device options. If the patients completing the AHP had preferences that were similar to those patients who participated in the CDRH study, we would expect that patients would assign a rating close to zero, although leaning some toward the end representing greater preference for a 30% reduction in total body weight. This evaluation would indicate that, although both attributes are important, a 30% reduction in total body weight is somewhat more important than a 1% risk of death due to the device. The process would then be repeated for all possible pairs of attributes or attribute levels, and statistical methods could then be used to estimate a full set of weights for the attributes or attribute levels.

4.1.6 Outranking Methods

Overview

Outranking methods compare a set of decision alternatives or profiles using pairwise comparisons. Unlike swing weighting and AHP, outranking methods base their comparisons on whether one profile is at least as good as or better than the other. The individual comparisons are then aggregated to develop a ranking of profiles in terms of preference. The ranking is ordinal and does not provide a set of weights measured on a common scale. However, outranking methods are commonly combined with direct weighting in which outranking is used to compare the alternatives and the direct weighting is used to elicit weights.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Panel approaches are typically used; survey approaches can be used
Are hypothetical scenarios required?	Panel methods typically used to evaluate real-world scenarios; survey methods can be used to evaluate hypothetical scenarios
How are attributes determined and defined?	Panel approaches typically use internal identification; survey approaches may require external identification
Is the method experimental?	Experimental design not required
Sample criteria	
What is the minimum sample size required?	No minimum sample size
What is the maximum sample size that can be reasonably achieved?	No maximum sample size for survey approaches; large sample sizes typically cost- and time-prohibitive for panel approaches
What is the time commitment required of patients?	Minutes to less than an hour required for survey approaches; time for getting to and attending the panel session required for panel approaches; preparation for panel session may also be required
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements for survey approaches; panel approaches may be more challenging; understanding of attributes and task required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used in certain circumstances
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used; specialized software available
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Panel approaches can allow for internal identification; survey approaches can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Possibly: When combined with direct weighting
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Possibly: When combined with direct weighting, ratios of weights could be interpreted as tradeoffs; however, tradeoffs are not elicited directly
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Possibly: Using survey approaches, differences in weights across individuals in a sample can be characterized; if panel approaches are used to achieve consensus among the panel, then this is not possible; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: Each patient would be presented with a set of device alternatives defined by attributes included in the study and asked to rank the alternatives from most preferred to least preferred. If the preferences of patients completing the outranking exercise are similar to those who participated in the CDRH study, we would expect that these patients would, on average, rank a profile including a 30% reduction in total body weight and a 1% risk of death due to the device as preferred to a profile including a 5% reduction in total body weight and a 0% risk of death due to the device. Direct weighting of the attributes would be required to elicit weights.

4.2 Health-State Utility Methods

Health-state utility methods yield an estimate of preferences for a health state (described as a single attribute or a profile) when compared with death and perfect health in which death is given a value of 0 and perfect health is given a value of 1. Higher weights equal greater preferences for a given health state. Health-state utility methods include TTO and SG.

4.2.1 Time Tradeoff

Overview

Time tradeoff is a stated-preference approach in which patients are asked to choose between living a specified time in a specified state of health and a shorter time in perfect health. Health states can be defined by individual attributes or by profiles. The time spent in perfect health then is varied until the patient is indifferent between longer life in the worse health state and the shorter life in perfect health. The ratio of the shorter amount of time in perfect health to the longer amount of time in the health state is the health-state utility. Higher values indicate greater preference for the health state. Time tradeoff health-state utilities can be compared directly to determine relative preferences for different health states or used as weights in models of incremental net benefits.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are required
Are hypothetical scenarios required?	Hypothetical scenarios are required
How are attributes determined and defined?	Can use internal or external identification
Is the method experimental?	Experimental design not required but could be used
Sample criteria	
What is the minimum sample size required?	Minimum sample size typically < 100
What is the maximum sample size that can be reasonably achieved?	No maximum sample size
What is the time commitment required of patients?	Minutes to less than an hour; total time requirement depends on number of scenarios to be evaluated
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements; understanding of attributes and tradeoff task required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Possibly: can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, weights for all attributes
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Possibly: ratios of weights could be interpreted as tradeoffs; however, tradeoffs are not elicited directly
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, differences in weights across individuals in a sample can be characterized; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: Using TTO may require a conceptual approach that is different than the one used by in the CDRH study. Such an approach would likely be similar to those approaches used in cost-utility analysis or incremental net-health benefit approaches in which TTO could be used to elicit preferences for outcomes (e.g., hospitalization requiring surgery), regardless of the probability or the duration of the outcome, and the weights are then applied to a series of health states over time in a probabilistic model. A description of these probabilistic modeling methods is beyond the scope of the Catalog; however, an example of this type of modeling approach is presented by Lynd et al. (45).

4.2.2 Standard Gamble

Overview

Standard gamble is a stated-preference approach in which patients are asked to choose between a certain outcome and a gamble between two uncertain outcomes, each with a probability of occurrence and where their aggregate chance of occurrence is 100%. Typically, the certain outcome is described as a health state. Health states can be defined by individual attributes or by profiles. The two outcomes in the gamble are death and perfect health. The probabilities of death and perfect health are varied until the patient is indifferent between the certain outcome and the gamble between death and perfect health. The probability of perfect health at which the patient is indifferent between the certain outcome and the gamble is the health-state utility.

Standard gamble also can be used to elicit risk tolerance directly. One minus the health utility can be interpreted as the maximum risk of death that would be tolerated in exchange for an improvement from experiencing the outcome to perfect health (i.e., elimination of the outcome or outcomes that define the health state).

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are required
Are hypothetical scenarios required?	Hypothetical scenarios are required
How are attributes determined and defined?	Can use internal or external identification
Is the method experimental?	Experimental design not required but could be used
Sample criteria	
What is the minimum sample size required?	Minimum sample size typically < 100
What is the maximum sample size that can be reasonably achieved?	No maximum sample size
What is the time commitment required of patients?	Minutes to less than an hour; total time requirement depends on number of scenarios to be evaluated
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements; understanding of attributes and tradeoff task required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used in certain circumstances
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used
Can the results be described and interpreted easily?	Methods and results easy to describe and interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Possibly: can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, weights for all attributes
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Possibly: ratios of weights could be interpreted as tradeoffs; however, tradeoffs are not elicited directly
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, differences in weights across individuals in a sample can be characterized; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: Using SG may require a conceptual approach that is different than the one used by in the CDRH study. Such an approach would likely be similar to those approaches use in cost-utility analysis or incremental net-health benefit approaches in which SG could be used to elicit preferences for outcomes (e.g., hospitalization requiring surgery), regardless of the probability or the duration of the outcome, and the weights are then applied to a series of health states over time in a probabilistic model. A description of these probabilistic modeling methods is beyond the scope of the Catalog; however, an example of this type of modeling approach is presented by Lynd et al. (45).

4.3 Stated-Preference Methods

Stated-preference methods are used to elicit relative preferences across attributes and changes in attribute levels using profiles. These methods differ from health-state utility methods because the weights elicited in these methods are not anchored on a cardinal scale between 0 and 1 in which 0 and 1 have a defined meaning. Instead, the weights elicited in stated-preference surveys can only be interpreted as ordinal rankings or relative preference weights. Stated-preference methods include direct assessment questions, threshold technique, conjoint analysis and discrete-choice experiments, and best-worst scaling.

4.3.1 Direct-Assessment Questions

Overview

Direct-assessment questions ask patients to provide a direct answer to a statement or relative-importance question. The statement or question asks the patient the extent to which he or she agrees with the statement regarding an attribute or profile or whether he or she prefers or would choose one attribute or profile over all other available attributes or profiles. When a statement is structured to evaluate the extent to which a patient values an attribute or the importance a patient places on an attribute, the result is a weight for that feature. Most direct assessments of profiles provide only a preferred profile or an ordinal ranking of profiles from the set of profiles presented to a patient and, thus, do not result in weights.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are required
Are hypothetical scenarios required?	Typically used to evaluate hypothetical scenarios; can be used to evaluate real-world scenarios
How are attributes determined and defined?	Typically use external identification; internal identification can be used
Is the method experimental?	Experimental design not required but could be used
Sample criteria	
What is the minimum sample size required?	No minimum sample size if experimental design not used
What is the maximum sample size that can be reasonably achieved?	No maximum sample size
What is the time commitment required of patients?	Minutes to less than an hour; total time requirement depends on number of scenarios to be evaluated
What are the cognitive and knowledge requirements of patients?	Minimal cognitive requirements; understanding of attributes required
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistics required; basic regression methods can be used in certain circumstances; more advanced analysis is possible
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used

Criteria	Review
Can the results be described and interpreted easily?	Results easy to describe and interpret. Basic methods are easy to describe. More advanced statistical methods may be more difficult to describe
Output criteria	
Can the method be used to identify attributes that are important to patients?	No
Can the method be used to estimate weights for attributes?	No
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	No
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, differences in choices across individuals in a sample can be characterized; differences in choices over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: Direct-assessment questions would simply require that patients be presented with alternative device profiles (see, for example, the set of profiles presented in Table 2 in Ho et al. [6]) and asked to choose among them. The proportion of patients preferring each profile can then be reported. If the patients completing the AHP have preferences that are similar to those patients who participated in the CDRH study, we would expect the results to look similar to those presented in the second column of Table 2 in Ho et al. (6). This method would need to be supplemented with additional patient-preference methods to determine the relative importance of different attributes or attribute levels to the choice of profile.

4.3.2 Threshold Technique

Overview

The threshold technique is a stated-preference approach in which patients are asked to choose between a reference profile and an alternative profile. Both the reference profile and the alternative profile are defined by a common set of attributes, although the levels of each attribute can vary between the two alternatives. In the threshold technique, one attribute is considered to be the study object. If the reference profile is chosen, the level of the study object in the alternative profile is improved until the patient changes his or her choice from the reference profile to the alternative profile. If the alternative profile is chosen, the study object in the alternative profile is made worse until the patient changes his or her choice from the alternative profile to the reference profile. The point at which the patient switches his or her choice is the threshold. If the attribute that is the study object is a risk, the threshold probability is an estimate of maximum acceptable risk. If the attribute that is the study object is an efficacy or effectiveness outcome (i.e., benefit), then the threshold probability is an estimate of minimum acceptable benefit.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are required
Are hypothetical scenarios required?	Hypothetical scenarios are required
How are attributes determined and defined?	Typically use external identification; internal identification can be used
Is the method experimental?	Experimental design not required, but could be used
Sample criteria	
What is the minimum sample size required?	No minimum sample size
What is the maximum sample size that can be reasonably achieved?	No maximum sample size
What is the time commitment required of patients?	Minutes to less than an hour; total time requirement depends on number of scenarios to be evaluated
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements; understanding of attributes and task required
Analysis criteria	
Does the method require statistical analysis?	Only descriptive statistic required; basic regression methods can be used in certain circumstances; more advanced analysis is possible

Criteria	Review
Does the method require specialized software?	Basic spreadsheet software required; commonly available statistical packages can be used
Can the results be described and interpreted easily?	Results easy to describe and interpret. Basic methods are easy to describe. More advanced statistical methods may be more difficult to describe
Output criteria	
Can the method be used to identify attributes that are important to patients?	No
Can the method be used to estimate weights for attributes?	No
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Yes, tradeoffs are a direct output of this method
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, differences in tradeoffs across individuals in a sample can be characterized; differences in weights over time can be characterized if data are elicited from the same sample at different points in time

Applying the method to the CDRH weight-loss example: The threshold technique requires specifying two profiles, similar to the profiles presented in Table 2 in Ho et al. (6). One profile is the reference profile and one profile is the alternative profile. One attribute in the alternative profile is then chosen as the attribute for which a threshold is to be estimated. If we assume that the reference profile includes a 5% reduction in total body weight and a 0% risk of death due to the device and the alternative profile includes a 30% reduction in body weight and a 1% risk of death due to the device, a patient would be asked to choose between these two profiles. If a patient chooses the alternative weight-loss device, then to estimate the maximum acceptable risk of death due to the device that a patient would be willing to accept to achieve the increase in total body weight loss from 5% to 30%, the rate of death due to the alternative device is increased incrementally until the patient prefers the reference profile. If a patient chooses the alternative device, then to estimate the minimum required increase in total body weight reduction, the percentage-point reduction in total body weight in the alternative profile is decreased incrementally until the person prefers the reference profile. If the patients completing the threshold-technique exercise have preferences that are similar to those patients who participated in the CDRH study, we would expect that the mean threshold for risk of death due to the device given a 25 percentage-point reduction in total body weight (30%-5%) would be approximately 1.3% and the mean threshold for increase in the percentage-point reduction in total body weight (above 5%) given a 1% risk of death would be approximately 22.4%. This process can be repeated for any attribute for which a threshold value is needed.

4.3.3 Conjoint Analysis and Discrete-Choice Experiments

Overview

In conjoint analysis and discrete-choice experiments, the attributes of each medical treatment are assigned different levels that can be combined into profiles, and the profiles are combined into groups of profiles known as *choice sets*. The profiles and choice sets are determined by an experimental design. Each patient is presented with a series of choice sets and asked to choose one profile in each choice set. Alternatively, a patient could be asked to rank profiles in a choice set or rate his or her strength of preference for one profile over an alternative profile or to allocate the percentage of patients that would be treated best with each alternative profile in each choice set. The pattern of responses is analyzed to estimate the rate at which patients are willing to trade off among the attributes and changes in attribute levels included in the study. The results can provide measures of the relative importance of attributes or changes in attribute levels and the rate of trade-off among attributes or attribute levels.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are required
Are hypothetical scenarios required?	Hypothetical scenarios are required
How are attributes determined and defined?	Typically use external identification; internal identification can be used
Is the method experimental?	Experimental design required
Sample criteria	
What is the minimum sample size required?	Minimum sample size typically 200-300
What is the maximum sample size that can be reasonably achieved?	No maximum sample size
What is the time commitment required of patients?	Typically 20 minutes to less than an hour; total time requirement depends on number of scenarios to be evaluated
What are the cognitive and knowledge requirements of patients?	Potentially significant cognitive requirements; understanding of attributes and trade-off task required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Advanced statistical analysis is typically required
Does the method require specialized software?	Analysis can be conducted using commonly available statistical packages; specialized software is also available
Can the results be described and interpreted easily?	Methods and results can be difficult to describe; advanced statistical methods make describing results difficult; numerous relative weights can make results difficult to interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Possibly: can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, weights for both attributes and changes in attribute levels
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Yes, tradeoffs among any attributes or changes in attribute levels
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, some statistical methods can provide quantitative estimates of the distribution of preferences across the sample; however, this method infers preferences from repeated observations from each patient; latent class analysis may be required to identify segments or subgroup analysis may be needed to test for differences in preferences; differences in preferences over time can be evaluated if the survey is implemented with the same patients at multiple points in time

4.3.4 Best-Worst Scaling

Overview

There are three types of best-worst scaling: object case, single-profile case, and multiple-profile case. In all cases, patients are presented with a set of alternatives and asked to identify the best or most important alternative and the worst or least important alternative. In the object case, attributes are combined into sets. Each set does not necessarily (and often does not) include all attributes. For each of a series of sets, patients are asked to indicate which of the attributes in the set is best or most desirable and which is worst or least desirable. In the single-profile case, each attribute takes on different levels. The attribute levels are combined into profiles. Patients are presented with a series of profiles and asked to indicate which attribute level in the profile is best or most desirable and which attribute level in the profile is worst or least desirable. In the multiple-profile case, attribute levels are combined into profiles, and the profiles are combined into sets of three or more. The multiple-profile case is very similar to a discrete-choice experiment. In each of a series of sets, patients are asked to indicate which profile is best or most desirable and which profile is worst or least desirable. In all three types of best-worst scaling, the pattern of responses is analyzed to estimate the relative importance of each attribute or attribute level.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Survey approaches are required
Are hypothetical scenarios required?	Hypothetical scenarios are required
How are attributes determined and defined?	Typically use external identification; internal identification can be used
Is the method experimental?	Experimental design required
Sample criteria	
What is the minimum sample size required?	No minimum sample size for simple methods. Minimum sample size required for regression methods
What is the maximum sample size that can be reasonably achieved?	No maximum sample size
What is the time commitment required of patients?	Minutes to less than an hour; total time requirement depends on number of scenarios to be evaluated
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements; understanding of attributes and task required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Simple analysis is possible; advanced statistical analysis is possible
Does the method require specialized software?	Basic spreadsheet software required for simple analysis; commonly available statistical packages can be used for more advanced analysis; specialized software packages are available
Can the results be described and interpreted easily?	Simple methods can be described easily; more advanced methods may be difficult to describe; results derived from simple methods may be more difficult to interpret because weights can be positive and negative; scaled results derived from more complex methods are easy to interpret
Output criteria	
Can the method be used to identify attributes that are important to patients?	Possibly: can allow for identification of important attributes from a set of externally identified attributes
Can the method be used to estimate weights for attributes?	Yes, weights for both attributes and changes in attribute levels Yes, tradeoffs among any attributes or changes in attribute levels
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	Yes, tradeoffs among any attributes or changes in attribute levels if attribute levels are included; ratios of attribute weights could be interpreted as tradeoffs if attribute levels are not included
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, some statistical methods can provide quantitative estimates of the distribution of preferences across the sample; however, this method infers preferences from repeated observations from each patient; latent class analysis may be required to identify segments or subgroup analysis may be needed to test for differences in preferences; differences in preferences over time can be evaluated if the survey is implemented with the same patients at multiple points in time

4.4 Revealed-Preference Methods

Revealed-preference methods are used to analyze patients' choices and behaviors in the real world. These methods can provide information on the number of patients for whom the benefits of a medical technology outweigh the risks and potentially the reasons why patients believe that benefits outweigh risks; however, unlike stated-preference methods, revealed-preference methods often cannot be used to derive weights for or the relative importance of individual attributes or changes in attribute levels. Revealed-preference methods include patient-preference trials and direct questions in clinical trials.

4.4.1 Patient-Preference Trials

Overview

Patient-preference trials are clinical trials in which patients are placed into arms of the trial depending on whether the patient has a strong preference for at least one of the medical technologies being studied. For example, in a trial with two medical technologies, A and B, patients with a strong preference for technology A are assigned to technology A, those with a strong preference for B are assigned to B, and those with no strong preference are randomly assigned to technologies A or B, effectively creating four study arms. An alternative approach is to randomize patients into two study arms. In the first arm, patients are randomly assigned to a medical technology. In the second arm, patients are assigned to a medical technology based on the patient's preference. If there are two technologies included in the study, then the study effectively has four arms. Follow-up questions can be administered to elicit the relative importance that each attribute of the medical technologies had on the patient's choice (or lack thereof).

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Clinical-study approaches are required
Are hypothetical scenarios required?	Real-world scenarios are required
How are attributes determined and defined?	External identification is required
Is the method experimental?	Experimental design required
Sample criteria	
What is the minimum sample size required?	Minimum sample size calculated based on expected effect sizes
What is the maximum sample size that can be reasonably achieved?	Large sample sizes typically cost- and time-prohibitive
What is the time commitment required of patients?	Significant time commitment associated with participation in a clinical trial
What are the cognitive and knowledge requirements of patients?	Minimal cognitive requirements; understanding of attributes required

Criteria	Review
Analysis criteria	
Does the method require statistical analysis?	Advanced statistical methods are required
Does the method require specialized software?	Commonly available statistical packages can be used
Can the results be described and interpreted easily?	Advanced statistical analysis may be more difficult to describe; results are easily interpreted
Output criteria	
Can the method be used to identify attributes that are important to patients?	No
Can the method be used to estimate weights for attributes?	No
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	No
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, differences in choices across individuals in a sample can be characterized; it may be impractical to implement the study with the same patients at different points in time

Applying the method to the CDRH weight-loss example: To evaluate patients' preferences for weight-loss devices using a patient-preference trial, patients could be randomized to those who choose which device to receive in the trial or to a randomization arm in which the device a patient receives is randomized. Patients who are allowed to choose the device they receive are explained the expected benefits and risks with each device and then directly asked which device they prefer.

4.4.2 Direct Questions in Clinical Trials

Overview

Direct questions consist of questions that ask patients in a clinical trial to indicate their choice between a medical technology with which they have had experience and either their current care or an alternative technology. Direct questions can be administered in any phase of clinical research. The most rigorous approach to administering direct questions is to conduct a phase 3 trial with a classic cross-over design in which each patient is exposed to first one medical technology and then another; the patient then is asked to indicate which technology he or she would like to or plans to continue at the conclusion of the study. Follow-up questions can then be administered to elicit the relative importance a patient's experience with each attribute of the medical technologies had on his or her choice.

Review

Criteria	Review
Methodology criteria	
How are the data acquired?	Clinical-study approaches are required
Are hypothetical scenarios required?	Real-world scenarios are required
How are attributes determined and defined?	External identification is required
Is the method experimental?	Experimental design required
Sample criteria	
What is the minimum sample size required?	Minimum sample size calculated based on expected effect sizes
What is the maximum sample size that can be reasonably achieved?	Large sample sizes typically cost- and time-prohibitive
What is the time commitment required of patients?	Significant time commitment associated with participation in a clinical trial
What are the cognitive and knowledge requirements of patients?	Some cognitive requirements; experience with attributes required
Analysis criteria	
Does the method require statistical analysis?	Advanced statistical methods are required
Does the method require specialized software?	Commonly available statistical packages can be used
Can the results be described and interpreted easily?	Advanced statistical analysis may be more difficult to describe; results are easily interpreted

Criteria	Review
Output criteria	
Can the method be used to identify attributes that are important to patients?	No
Can the method be used to estimate weights for attributes?	No
Can the method be used to estimate the tradeoffs that patients are willing to make among attributes?	No
Can the method be used to detect, describe, or quantify heterogeneity in preferences across patients and across time?	Yes, differences in choices across individuals in a sample can be characterized; it may be impractical to implement the study with the same patients at different points in time

Applying the method to the CDRH weight-loss example: Including direct questions in clinical trials of weight-loss devices may not be possible because most clinical trials that include direct preference questions require that patients experience both options. This may not be possible in the case of implantable weight-loss devices. Temporary devices implanted endoscopically or nonsurgical weight-loss devices may allow for the type of cross-over design required to use direct questions to elicit preferences in clinical trials. In such examples, the patient would be asked to state which technology they would continue to use after experiencing the benefits and being exposed to the risks of each device.

5 EXAMPLES OF PRIOR USE OF METHODS

In this section, we review some examples of prior use of the methods in the Catalog (structured weighting, health-state utilities, stated preference, and revealed preference) in benefit-risk assessments. The remainder of this section considers examples in each of the categories. Definitions of methods are repeated for those specific methods for which examples of prior use have been identified.

Hauber et al. (20) and Mt-Isa et al. (46) have described the use of many of these methods in benefit-risk assessment and provided numerous empirical examples. The Innovative Medicines Initiative (IMI) Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) project resulted in a number of case studies of benefit-risk assessments in which preferences were incorporated. Recommendations from IMI-PROTECT are presented in Hughes et al. (47). In the Catalog, we identify the case studies evaluated by the IMI-PROTECT working group as examples of prior use where appropriate and provide citations to allow users to review the relevant case-study reports. We do not describe each of the IMI-PROTECT case studies or empirical examples identified by Hauber et al. (20) and Mt-Isa et al. (46) in detail. Instead, we focus on identifying additional examples of the prior use of each method in benefit-risk analysis when they exist. For some methods included in the Catalog, no examples of prior use of the method in eliciting patients' benefit-risk preferences exist to the best of our knowledge.

5.1 Examples of Structured-Weighting Studies

Structured-weighting methods included in the Catalog are simple direct weighting, ranking exercises, swing weighting, point allocation, AHP, and outranking. Simple direct weighting, ranking exercises, swing weighting and point allocation are methods for eliciting the relative importance of benefit and risk outcomes and typically are used as part of a decision analysis such as MCDA. Analytic hierarchy process and outranking methods are decision-analysis methods that include both an assessment of the magnitude of relevant benefit and risk outcomes and the importance of benefit and risk outcomes. Decision-analysis methods have been used in benefit-risk analyses for regulatory decisions; however, most of these applications have been conducted using expertise and judgment of clinical experts or other professionals rather than patients (see Levitan et al. [48]).

Ranking is a method for placing a set of attributes, attribute levels, or profiles in order of increasing or decreasing preference or importance. Ranking may be strict (no ties permitted) or nonstrict (ties permitted). Often, no numeric values reflecting the magnitude of preference are provided. Point allocation requires that each feature attribute in a set be assigned points proportional to the importance associated with specified changes in each

feature attribute from its lowest level to its highest level. The total number of points to be allocated among the attributes is fixed. The resulting values are usually normalized and interpreted as weights for the set of attributes. Higher weights indicate greater importance.

- Stafinski et al. (8) used multiple methods, including ranking and point allocation, to elicit the relative importance of different cardiovascular outcomes to patients with coronary disease or previous MI. Each patient was asked to complete multiple exercises to evaluate the relative importance of cardiovascular outcomes, including death, cardiogenic shock, congestive heart failure, and repeat MI. The ranking exercise asked patients to rank these events from most severe to least severe. The proportion of patients assigning the same rank to each outcome was calculated. In the point-allocation exercise, patients were asked to allocate 20 points among the four endpoints with more points indicating greater severity. The mean number of points allocated to each outcome was reported. In both exercises, death was considered worse than cardiogenic shock, which, in turn, was considered worse than congestive heart failure. Repeat MI was the least important among the four outcomes.

Swing weighting requires that each attribute in a set be assigned a range of minimum to maximum levels, such that the full range of expected levels is included within the range. The attributes are then ranked in decreasing order of the importance that a change in each attribute, from its lowest level to its highest level, would have on a given decision. The attribute with the highest rank is assigned a weight of 100. The second attribute is then assigned a weight on a scale from 1 to 100, reflecting the degree a swing from its lowest to highest level would influence the decision, compared with the highest-ranked feature. Thus, higher weights indicate greater importance. This process is repeated for all attributes. The resulting weights are normalized to sum 100 and provide a weight for each attribute over the range of levels assigned to that feature.

- The IMI-PROTECT Benefit-Risk Group conducted six benefit-risk case studies in which swing weighting was used as part of an MCDA (49-54).

Analytic hierarchy process requires that changes in the levels of each attribute are compared to derive weights that indicate the relative importance of changes in attribute levels to achieving a decision goal. This is accomplished through a series of pairwise comparisons between every pair of attributes. The questions are used to indicate which attribute is preferred, more likely, or more important as well as the strength of preference. Software interrogates a patient when choices are contradictory. These comparisons then are used to compute a weight for each attribute. For beneficial attributes, higher weights indicate greater preference. For undesirable attributes, higher weights indicate lower preference.

- Hummel et al. (7) provided an example of how AHP can be used to conduct a benefit-risk assessment. The AHP was illustrated using a hypothetical example of tissue regeneration for repairing small cartilage lesions in the knee. The benefits and risks considered in this example included effectiveness, adverse events, and surgical procedure. These researchers concluded that the increased benefits of tissue-engineered cartilage exceeded the increased risks of adverse events and the increased burden of surgery when compared with standard of care.

5.2 Examples of Health-State Utility Studies

Health-state utility methods include SG and TTO. Both SG and TTO can be used to provide measures of relative preference for benefit and risk outcomes. Standard gamble is a stated-preference approach in which patients are asked to choose between a certain outcome and a gamble between two uncertain outcomes, each with a probability of occurrence and where their aggregate chance of occurrence is 100%. The probabilities of the uncertain outcomes are varied until the patient is indifferent between the certain outcome and the gamble between the alternatives. Conventionally, SG is used to elicit health-state utilities for use in cost-utility models. In studies using SG to estimate health-state utilities, the certain health state is the health state for which the utility is estimated, and the uncertain health states are death and perfect health. Standard gamble health-state utilities can be used as weights in models of incremental net benefits. Standard gamble also can be used to elicit risk tolerance directly.

Time tradeoff is a stated-preference approach in which patients are asked to choose between living a specified time in a health state and a shorter time in perfect health. The time in perfect health is varied until the patient is indifferent between longer life in the worse health state and the shorter life in perfect health. Time tradeoff health-state utilities can be used as weights in models of incremental net benefits. Two examples of the use of health-state utilities in benefit-risk analysis are described below. In addition, Hauber et al. (20) described an example of the use of SG in benefit-risk analysis proposed by O'Brien et al. (17).

- Lynd et al. (45) conducted a benefit-risk analysis of rofecoxib relative to naproxen to treat arthritis. The research team used TTO-based health-state utility estimates from an existing cost-effectiveness analysis. Risks in the study included gastrointestinal bleeding, gastrointestinal perforation, dyspepsia, acute MI, and the risk of death. The study team developed a discrete-event simulation model to model benefit and risk outcomes over a 1-year time horizon using data from clinical trials. Health-state utility weights were applied to the corresponding outcomes. The model was used to calculate incremental gains in quality-adjusted life-years resulting from using rofecoxib instead of naproxen. Gains in quality-adjusted life-years were positive

(favoring rofecoxib) in 94% of the model simulations. The researchers concluded that the benefits or rofecoxib are likely to exceed the risks.

- Johnson & Johnson Pharmaceutical Research & Development (55) conducted a weighted quantitative benefit-risk analysis of rivaroxaban for the prevention of deep vein thrombosis and pulmonary embolism in patients undergoing hip or knee replacement surgery using health-state utilities. The study team reviewed utilities from numerous existing sources and developed three health-state utility estimates (low, typical, high) for each potential outcome. They then applied the change in health-state utility to the number of excess events for each type of event using pooled data from clinical trials to calculate the net utility of rivaroxaban versus enoxaparin. The authors demonstrated that an increase in utility indicated that the net benefits of rivaroxaban outweighed the net risks in this indication.

5.3 Examples of Stated-Preference Studies

Stated-preference methods include direct-assessment questions, threshold technique, conjoint analysis and discrete-choice experiments, and best-worst scaling.

Direct-assessment questions ask patients to provide a direct answer to a statement or relative-importance question. The statement or question asks the patient the extent to which he or she agrees with the statement regarding an attribute or profile or whether he or she prefers or would choose one attribute or profile over all other available attributes or profiles. When a statement is structured to evaluate the extent to which a patient values an attribute or the importance a patient places on an attribute, the result is a weight for that feature. Most direct assessments of profiles provide only a preferred profile or an ordinal ranking of profiles from the set of profiles presented to a patient and, thus, do not result in weights.

- Sarkissian et al. (18) presented patients with three options for the management of asymptomatic renal calculi: a surgical option, shock wave therapy, and observation. Each option was defined by benefit and risk attributes. Patient choices among these three options were then regressed on patient characteristics. Patients were also given the opportunity to defer the treatment decision to their physicians. These authors concluded that patients' choice of treatment was influenced by their prior treatment experiences and that most patients preferred to allow the physician to make the choice.

- Yachimski et al. (19) presented patients with two options for the treatment of nondysplastic Barrett's esophagus: endoscopic ablation or aspirin. Each option was defined by benefit and risk attributes. In addition, the level of endoscopic surveillance was varied systematically for both options. The authors found that most patients preferred endoscopic ablation regardless of the frequency of surveillance. In addition, the authors found no correlation between patients' demographic characteristics of health history and treatment choice.

The threshold technique is a stated-preference approach in which patients are asked to choose between a reference treatment and an alternative treatment. Both the reference treatment and the alternative treatment are defined by a common set of treatment attributes, although the levels of each attribute can vary between the two alternatives. In the threshold technique, one attribute is considered to be the study object. If the reference treatment is chosen, the study object in the alternative treatment is improved until the patient changes his or her choice from the reference treatment to the alternative treatment. If the alternative treatment is chosen, the study object in the alternative treatment is made worse until the patient changes his or her choice from the alternative treatment to the reference treatment. The point at which the patient switches his or her choice is the threshold. Hauber et al. (20) described numerous examples of the use of threshold techniques in benefit-risk analysis. Three additional examples are described as follows:

- Stafinski et al. (8) used a threshold technique to evaluate willingness to accept risks of systemic bleed and nonfatal intracranial hemorrhage in exchange for reducing the risk of death, cardiogenic shock, congestive heart failure, or repeat MI among patients with coronary disease or previous MI. The authors calculated the proportion of patients who would accept varying levels of risks of systemic bleed and nonfatal intracranial hemorrhage for each of multiple scenarios. The results of this study indicate that most patients are willing to accept increases in the risk of systemic bleed and intracranial hemorrhage in exchange for significant decreases in the risks of cardiovascular outcomes.
- Kennedy et al. (21) used a threshold technique to evaluate the minimum required increase in treatment effect for five pairwise comparisons of treatments for Crohn's disease. Each treatment-choice question was followed by a question in which patients were asked to choose a reason that best described why they chose one treatment over another. These authors found significant heterogeneity among patients in the sample in the minimum required benefit thresholds. This heterogeneity was not explained by differences in patients' demographic characteristics or health history.

- Kok et al. (22) assessed expectant parents' (both mothers' and fathers') preferences for vaginal or cesarean delivery of a fetus in breech presentation. The study was designed to estimate the threshold for risks of neonatal complications at which patients would switch their choice of delivery. Patients were then asked to rate the importance of each attribute of the delivery options using a Likert scale. Most patients preferred cesarean delivery to vaginal delivery but were sensitive to changes in neonatal complication risks. The risk of neonatal complications at 2 years was the most important attribute for mothers. The health of the mother was the most important attribute to fathers.

Discrete-choice experiments are a form of conjoint analysis in which a medical technology is decomposed into a set of attributes. Each of the attributes is assigned different levels. The attribute levels are combined into profiles, and the profiles are combined into choice sets according to an experimental design. Each patient is presented with a series of choice sets and asked to indicate which profile he or she would choose in each of a series of choice sets. The pattern of responses is analyzed to estimate the rate at which patients are willing to trade off among the attributes included in the study. Hauber et al. (20) provided numerous examples of the use of discrete-choice experiments in benefit-risk analysis. A more recent example is described as follows:

- In 2011, the CDRH commissioned a pilot study to conduct a discrete-choice experiment to elicit benefit-risk preferences for the attributes of weight-loss technologies among Americans with a body mass index of 30 kg/m^2 or greater (6). The discrete-choice experiment was designed to measure tradeoffs that patients were willing to make among total weight loss, duration of weight loss, duration of mild-to-moderate side effects, mortality risk, risk of a side effect requiring hospitalization, recommended dietary restrictions, reduction in risk of comorbidity or reduction in prescription dosage for existing comorbidity, and type of surgery. The results of this study included maximum risk tolerance for each risk for different levels of weight loss and the minimum weight loss required to accept different levels of each risk. CDRH is using the study tool to define minimum clinical effectiveness for a given technology profile and evaluate new weight-loss technologies.

Additional examples of discrete-choice experiment studies include Mühlbacher and Bethge (23), Fraenkel et al. (24), Wouters et al. (25), and Guimaraes et al. (26).

There are three types of best-worst scaling: object case, single-profile case, and multiple-profile case. In the object case, the attributes are combined into sets. For each of a series of sets, patients are asked to indicate which of the attributes is best or most desirable and which is worst or least desirable. In the single-profile case, each attribute can take on different levels. The attribute levels are combined into profiles. Patients are presented with a series of profiles and asked to indicate which attribute level is best or most desirable and

which attribute level is worst or least desirable for each profile. In the multiple-profile case, attribute levels are combined into profiles, and the profiles are combined into sets of three or more. In each of a series of sets, patients are asked to indicate which profile is best or most desirable and which profile is worst or least desirable. In all three types of best-worst scaling, the pattern of responses is analyzed to estimate the relative importance of each attribute or attribute level.

- Yuan et al. (4) used object-case best-worst scaling to elicit the relative importance of different cardiovascular outcomes to patients with acute coronary syndrome and to cardiologists. Patients and physicians were asked to choose the most concerning and least concerning outcomes in a series of questions. Each question included a subset of outcomes from the full set of possible outcomes. Possible outcomes included death, various levels of stroke, MI, and bleeding. The relative importance of each outcome was scaled relative to death so that the importance of each outcome could be interpreted as the number of deaths that would be equivalent to one case of that outcome. Among patients and physicians, nonfatal disabling stroke was viewed as equivalent to or worse than death, and all levels of bleeding were viewed as significantly less concerning than death. The results of this study provided a complete set of importance weights that could be used to evaluate the benefits and risks of antithrombotic medical technologies for acute coronary syndrome.
- Peay et al. (27) conducted a study among caregivers of patients with Duchenne muscular dystrophy using object-case best-worst scaling. Each caregiver was presented with a series of questions, each including six outcomes from a set of 18 benefit and risk outcomes. The outcomes included varying levels of a treatment's effect on muscle function, life expectancy, knowledge about the treatment, nausea, risk of bleeds, and risk of arrhythmia. The study team found the treatment's effect on muscle function was the most important feature, followed by the risk of arrhythmia and the risk of bleeding. These researchers concluded that caregivers were willing to accept treatment risks to improve muscle function even if the treatment did not increase life expectancy.

5.4 Examples of Revealed-Preference Studies

Revealed-preference methods for benefit-risk analysis include patient-preference trials and direct questions in clinical trials. Patient-preference trials are clinical trials in which patients are placed into arms of the trial depending on whether the patients have a strong preference for at least one of the treatments being studied. For example, in a trial with two treatments, A and B, patients with a strong preference for treatment A are assigned to treatment A, those with a strong preference for B are assigned to B, and those with no strong preference are randomly assigned to treatments A or B, effectively creating four treatment arms. These studies are of value when patients (or clinicians) have such strong treatment preferences that they refuse randomization. The absence of these patients from trials may restrict generalization of the results, as patients may not be representative. A further potential source of bias exists when patients with strong treatment preferences are recruited and randomized. When it is not possible to blind patients to their treatment allocation, as often occurs with medical technology trials, they may experience resentful demoralization if they do not receive their preferred treatment, and they may have poor treatment compliance. On the other hand, patients receiving their preferred treatment may have better than average treatment compliance. Therefore, there may be a treatment effect that results from patient preferences and not from therapeutic efficacy.

- Torgerson and Sibbald (56) described a perceived need for patient-preference trials. Tilbrook et al. (57) conducted a systematic review of patient-preference trials in musculoskeletal diseases and determined that patient preferences were systematically correlated with outcomes and persistence in these trials. Marcus et al. (58) demonstrated analytically how to derive an unbiased estimate of the effect of randomization versus preferences on endpoints.
- Crowther et al. (28) conducted a patient-preference trial of outcomes of vaginal birth versus cesarean among women with prior cesarean. Women were assigned to a randomization arm or a preference arm. The vast majority were assigned to the preference arm with only a very small number assigned to the randomization arm. The authors found that the risk of fetal or infant death or serious adverse infant outcome was lower among those with planned elective cesarean when compared with planned vaginal birth but that the risk of major maternal hemorrhage was greater among the cesarean group. The authors did not attempt to compare the results between the randomization arm and the patient-preference arm of the trial.

- Long et al. (29) described a patient-preference trial of behavioral interventions for women with heart disease. Women were assigned to a randomization arm or a patient-preference arm. The results of the study indicated that women in the patient-preference arm had greater improvements in sickness-impact-profile scores and were more likely to adhere to the behavioral intervention than women in the randomization arm. Through this study, the authors described methods for estimating the causal effect of patient preferences on improvements in outcomes.

Direct questions in clinical trials consist of questions that ask patients in a clinical trial to indicate their choice between a medical technology with which they have had experience and either their current care or an alternative technology. Direct questions can be administered in any phase of clinical research. The most rigorous approach to administering direct questions may be to conduct a phase 3 trial with a classic cross-over design in which each patient is exposed to first one medical technology and then another; the patient then is asked to indicate which technology he or she would like to continue with at the conclusion of the study.

- Escudier et al. (30) reported the results of a cross-over trial of sunitinib and pazopanib for the treatment of renal cell carcinoma. Patients in the trial were exposed to both sunitinib and pazopanib and then asked which of the two drugs they preferred. A majority of patients preferred pazopanib. A minority of patients preferred sunitinib or had no preference. Additional questions were included at the conclusion of the trial in an effort to determine which factors most influenced patients' preferences. However, it is important to note that patients were not informed of the efficacy of either drug when asked to indicate their preferences. Therefore, preferences were based on patients' experiences with adverse events and dosing schedule.

6 AREAS FOR FUTURE RESEARCH REGARDING PATIENT-PREFERENCE METHODS

The Catalog was reviewed by a number of individuals at different stages during its development. These reviewers included members of the PCBR project Steering Committee, CDRH staff, and MDIC member companies. The feedback from these reviewers included questions for which no clear answers or guidance currently exists. These questions relate to the use of patient-preference methods in general and provide some indication of areas for future research into patient-preference methods that may increase the transparency, validity, and ultimate utility of patient-preference studies in benefit-risk assessments. The Working Group sought to identify the issues about which the feedback was concerned and provide some suggestions for future research. The list of issues outlined below is not intended to be exhaustive but to reflect themes that emerged during the development of the Catalog and questions provided by reviewers in response to earlier versions of the Catalog. In addition, although there may be many areas in which further research would provide a greater understanding of properties of individual methods, the questions listed here are intended to apply generally to the use of all patient-preference methods in benefit-risk analysis and not to address the potential lack of experience with an individual method.

The list of issues that may be answered by future research falls into four broad categories: the choice of method, the sample, the design development of a study, and the validity of the method. One or more suggestions for future research are provided to address questions. The suggestions for future research are only suggestions and are not meant to be prescriptive. There may be other approaches to providing information to address these questions.

6.1 Choice of Method

The question posed most often by reviewers was how to choose a specific method in any individual situation. At the present, there is no algorithmic answer to this question. As described in Section V of the Framework, different types of information probably are required at different stages of the product lifecycle. In addition, different methods provide different types of information. Therefore, several methods could be relevant at different stages of the product lifecycle. In addition, the Catalog describes the ability of methods to provide different types of outputs. However, beyond this information, little is known about how well any given patient-preference method performs relative to other potential methods in any given situation.

Issue: Once a regulator or sponsor has identified the type of patient-preference information required and identified a set of methods that could potentially provide that information, the best option among this set of alternative methods may be difficult to determine. Another way to think about this question is whether using different methods to answer the same research question will yield results sufficiently consistent to lead to the same decision.

Suggestion for Future Research: There are only a few studies comparing the outputs of different patient-preference methods applied to the same underlying research question (8, 39-41). We recommend that additional studies be conducted in which multiple patient-preference methods are used to address the same research question. Such studies will enable users to directly compare and contrast the performance of different patient-preference methods and their implications for decision making, along with the relative advantages and limitations of any given method in a different situation.

6.2 Sample

As described in the Catalog, there are no clear guidelines for determining the sample from which patient-preference data should be derived. There are a number of challenges to determining the appropriate sample. First, there is no consensus on the extent to which samples used in patient-preference studies should be representative of a larger population. The need for sample representativeness can vary across a product lifecycle and with the specific research question of interest. Finally, there are several potential ways to test for, explain, and/or control for heterogeneity of preferences within a sample or across a population; however, these methods are applied inconsistently, if at all, in the literature.

Issue: There is no clear guidance on whose preferences should be measured in a patient-preference study. Representativeness is evaluated entirely by comparing the sample with the population of interest; thus, the representativeness of any sample will be determined, in part, by the research question. The research question may involve understanding the preferences of a population with well-defined characteristics (e.g., a clinical trial population for which there are well-defined inclusion and exclusion criteria). In this case, recruiting a representative sample is relatively straightforward. However, the research question may be broader and involve understanding the preferences of the population that will be exposed to the medical technology in the future. In this case, recruiting a representative sample may be difficult because the characteristics of the overall population of interest may not be well understood. For example, a medical technology may be indicated to treat a given condition, but the number of patients with that condition and the distribution of ages and genders of patients in that population may not be known with any degree of certainty. Even if observable characteristics of the

population are known, it is impossible to ensure that the preferences of any sample are representative of the overall population because differences in preferences may not be completely explainable by observable characteristics.

> **Suggestion for Future Research:** Because it is difficult to know how the representativeness of a sample is likely to affect the results of a patient-preference study, it may be important to conduct the same patient-preference study with different samples with different characteristics. Such a study would provide evidence regarding the sensitivity of the results obtained with specific patient-preference methods to the choice of sample and may provide evidence of systematic biases resulting from sampling choice.

Issue: Patient preferences for the benefits and harms of medical technologies are likely not consistent within populations. Sometimes variations in preferences can be attributed to differences in observable characteristics of sampled patients such as age, weight, and sex and diagnostic variables. Of particular interest in benefit-risk assessments of medical technologies is whether people with prior experience with the medical technology or a similar technology have preferences that vary systematically from those people who do not have such experience. Whether prior experience influences patients' preferences for medical technologies in any case, every case, or only in cases with certain properties is unknown.

> **Suggestion for Future Research:** A single study of patients' benefit-risk preferences for medical technologies might be conducted with samples of patients who have prior experience with a medical technology as well as patients who may potentially be eligible for a medical technology to provide evidence of the extent to which people with prior experience have systematically different preferences from those who do not. This type of study might be repeated for different types of medical technologies to provide evidence regarding the extent to which such differences in preferences may or may not exist for different technologies.
>
> Alternatively, a patient-preference study could be conducted among patients who would be eligible for a medical technology but who have no prior experience with the medical technology. The same patient-preference study could then be conducted among patients who receive the medical technology once the medical technology becomes available. Information on differences in preferences between these groups would provide both an understanding of the effect of an experience on patient preferences and may also provide a method for validating the premarket patient-preference study (see suggestion for future research under Study Validity).

6.3 Development of the Study

The first step in conducting a benefit-risk preference study is to define the research question based on the study objective. Study objectives can span the range from eliciting patient views on a medical technology for guiding development decisions to eliciting rigorous statistical evidence of benefit-risk tradeoffs from a sample of patients to provide evidence for regulatory decision making. Implicit in the research question is the selection of a set of attributes of the medical technology to be evaluated in the patient-preference study.

Issue: There is no definitive guidance on the selection of attributes of medical technologies to be used in a preference study, yet the choice of attributes is critical. Different methods have been used to determine which attributes should be included in a patient-preference study. For some studies, the primary objective is to identify the attributes of the medical technology that are important to patients. However, in other studies, the objective is to quantify the relative importance of attributes or to quantify the tradeoffs patients are willing to make among these attributes. A variety of methods exists to identify attributes. Sometimes the attributes that are identified as important by patients during qualitative research are used in quantitative studies. Other approaches to identifying the attributes for a study include asking a group of medical or regulatory experts to identify those attributes that are most important to a regulatory decision or conducting a literature review or review of product labels to determine those attributes that distinguish one medical technology from alternative medical technologies or a standard of care. Benefit-risk frameworks, such as the Benefit-Risk Action Team (BRAT) Framework and PrOACT-URL, also provide guidance on attribute selection (59-63). However, these methods are applied inconsistently across patient-preference methods and across patient-preference studies using a given method.

Suggestion for Future Research: A patient-preference study designed to determine the impact of changing the list of attributes with any given method could be conducted. Such a study could have two arms in which patients are assigned randomly to see different sets of attributes developed using different approaches to attribute identification. A key component of such a study would be to ensure that a number of attributes (perhaps half) are common to both studies. The results of a study such as this could provide an understanding of whether differences in attribute selection result in comparable weights or tradeoffs for a common set of attributes or whether the inclusion of different attributes affects patients' reviews of the common set of attributes.

Issue: In addition to determining the attributes included in a survey-based patient-preference study, researchers must define those attributes for patients, and there is no definitive guidance on how to accomplish this. Differences in the definition of an attribute could lead to different estimates of relative importance or tradeoffs. In addition, different levels of comprehension of an attribute definition could lead to different reviews of those attributes. The extent to which differences in attribute definitions and patients' comprehension of those definitions affect patient-preference estimates.

Suggestion for Future Research: A patient-preference study designed to determine the impact of changing definitions of attributes could be conducted. Such a study could include two arms in which patients are assigned randomly to one of two sets of attribute definitions could be conducted. The results of a study such as this could provide an understanding of the extent to which patient-preference estimates are sensitive to the way in which attributes are described. In addition, comprehension questions could be included to evaluate patients' understanding of the definitions presented in the survey, and the data could be analyzed to test whether differences in the level of comprehension systematically affect patient-preference estimates.

6.4 Study Validity

Stated-preference methods typically involve scenarios in which patients are asked to make hypothetical choices without actually experiencing the consequences of that choice. Therefore, it is unknown whether patients would actually do what they say they would do. Often hypothetical choices are necessary because observing actual choices is impossible or observing actual choices does not provide sufficient variation in attributes and attribute levels to tease out the rates at which patients would be willing to trade off among attributes. Despite providing experimental control over the attributes and attribute levels that are considered in treatment decisions, the hypothetical nature of the choice from which most patient-preference data are derived may undermine the validity of patient-preference estimates.

Issue: There is no clear definition of what constitutes a valid patient-preference study. Unlike in PRO research, there is not a standard set of validity tests that can be applied to patient-preference studies. In addition, it is not yet clear what regulators or other users of patient-preference data would need to be comfortable with using results from a patient-preference study.

Suggestion for Future Research: A review of standards and methods for assuring validity in other types of clinical studies, such as studies using PROs, might identify principles that could be useful in developing analogous, but likely different, approaches to validating patient preference studies.

Issue: Although some methods exist for evaluating the consistency of hypothetical choices when patients are asked to make multiple choices in the same study, there is little evidence as to what level of consistency would be required for a study to be considered valid. One method for establishing the validity of patient-preference methods is to determine whether the hypothetical choices patients make are consistent across choice scenarios and whether these choices are consistent with actual choices made later after exposure to a therapy.

Suggestion for Future Research: A review of existing patient-preference studies to examine the consistency of responses using hypothetical choices might provide some guidance as to the level of consistency that can be expected from such studies.

Issue: As noted above, stated-preference methods typically involve scenarios in which patients are asked to make hypothetical choices, and it is unknown whether patients would actually do what they say they would do.

Suggestion for Future Research: A patient-preference study designed to generate information that can help validate the evidence collected through premarket patient-preference studies and to understand the effect of experience on patient preferences could be conducted (see suggestion for future research under Sample). One study could be conducted among patients who would be eligible for a medical technology but who have no prior experience with the medical technology. The same patient-preference study could then be conducted among patients who receive the medical technology once the medical technology becomes available.

7 CITATIONS

1. US Food and Drug Administration (FDA). Guidance for industry and Food and Drug Administration staff—factors to consider when making benefit-risk determinations in medical device premarket approvals and de novo classifications. March 28, 2012. Available at: http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm267829.htm. Accessed August 25, 2014.

2. Fraenkel L, Seng EK, Cunningham M, Mattocks K. Understanding how patients (vs physicians) approach the decision to escalate treatment: a proposed conceptual model. Rheumatology. 2015;54:278-85.

3. Fried TR, Tinetti ME, Towle V, O'Leary JR, Iannone L. Effects of benefits and harms on older persons' willingness to take medication for primary cardiovascular prevention. Arch Internal Med. 2011;171(10):923-8.

4. Yuan Z, Levitan B, Burton P, Poulos C, Hauber AB, Berlin JA. Relative importance of benefits and risks associated with antithrombotic therapies for acute coronary syndrome: patient and physician perspectives. Curr Med Res Opin. 2014 2014;30(9):1733-41.

5. Wilson L, Loucks A, Bui C, Gipson G, Zhong L, Schwartzburg A, et al. Patient centered decision making: use of conjoint analysis to determine risk-benefit trade-offs for preference sensitive treatment choices. J Neurol Sci. 2014;344(1-2):80-7.

6. Ho MP, Gonzalez JM, Lerner HP, Neuland CY, Whang JM, …, Hauber AB, et al. Incorporating patient-preference evidence into regulatory decision making. Surg Endosc. 2015 Jan 1. [Epub ahead of print]

7. Hummel JM, Bridges JF, IJzerman MJ. Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. Patient. 2014;7(2):129-40.

8. Stafinski T, Menon D, Nardelli A, Bakal J, Ezekowitz J, Tymchak W, et al. Incorporating patient preferences into clinical trial design: results of the opinions of patients on treatment implications of new studies (OPTIONS) project. Am Heart J. 2015;169(1):122-31.

9. Edwards W, Barron FH. SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. Organizational Behavior and Human Decision Processes. 1994;60 (3):306-25.

10. Barron FH, Barrett, BE. The efficacy of SMARTER—Simple multi-attribute rating technique extended to ranking. Acta Psychologica. 1996;93(1):23-36.

11. McCaffrey J, Koski N. Competitive analysis using MAGIQ. Msdn magazine, 2006 (October). Available at: http://msdn.microsoft.com/en-us/magazine/cc135436.aspx. Accessed November 20, 2014.

12. Caster O, Norén GN, Ekenberg L, Edwards IR. Quantitative benefit-risk assessment using only qualitative information on utilities. Med Decis Making. 2012;32(6):E1-15.

13. Dolan JG. Shared decision-making – transferring research into practice: the Analytic Hierarchy Process (AHP). Patient Educ Couns. 2008;73(3):418-25.

14. Singh S, Dolan JG, Centor RM. Optimal management of adults with pharyngitis – a multi-criteria decision analysis. BMC Med Inform Decis Mak. 2006;6(14): doi:10.1186/1472-6947-6-14.

15. Avila M, Becerra V, Guedea F, Suarez JF, Fernandez P, Macias V. Estimating preferences for treatments in patients with localized prostate cancer. Int J Radiat Oncol. 2015;91(2):277-87.

16. Kuchuk I, Bouganim N, Beusterien K, Grinspan J, Vandermeer L, Gertler S, et al. Preference weights for chemotherapy side effects from the perspective of women with breast cancer. Breast Cancer Res Treat. 2013;142:101-7.

17. O'Brien BJ, Elswood J, Calin A. Willingness to accept risk in the treatment of rheumatic disease. J Epidemiol Community Health. 1990;44(3):249-52.

18. Sarkissian C, Noble M, Li J, Monga M. Patient decision making for asymptomatic renal calculi: balancing benefit and risk. Urology. 2013;81:236-40.

19. Yachimski P, Wanu S, Givens T, Howard E, Higginbotham T, Price A, et al. Preference of endoscopic ablation over medical prevention of esophageal adenocarcinoma by patients with Barrett's esophagus. Clin Gastroenterol Hepatol. 2015;13:84-90.

20. Hauber AB, Fairchild A, 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.

21. Kennedy ED, To T, Steinhart AH, Detsky A, Llewellyn-Thomas HA, McLeod RS. Do patients consider postoperative maintenance therapy for Crohn's disease worthwhile? Inflamm Bowel Dis. 2008;14(2):224-35.

22. Kok M, Gravendeel L, Opmeer BC, van der Post JAM, Mol BWJ. Expectant patents' preferences for mode of delivery and trafe-offs of outcomes for breech presentation. Patient Education and Counseling. 2008;72:305-410.

23. Mühlbacher AC, Bethge S. Reduce mortality risk above all else: a discrete-choice experiment in acute coronary syndrome patients. Pharmacoeconomics. 2015;33:71-81.

24. Fraenkel L, Suter L, Cunningham C, Hawker G. Understanding preferences for disease-modifying drugs in osteoarthritis. Arthritis Care Res. 2014;66(8):1186-92.

25. Wouters H, Maatman GA, van Dijk L, Bouvy ML, Vree R, can Geffen ECG, et al. Trade-off preferences regarding adjuvant endocrine therapy among women with estrogen receptor-positive breast cancer. Ann Oncol. 2013;24:2324-9.

26. Guimaraes C, Marra CA, Gill S, Simpson S, Meneilly G, Queiroz RHC, Lynd LD. A discrete choice experiment review of patients' preferences for different risk, benefit, and delivery attributes of insulin therapy for diabetes management. Patient Prefer Adherence. 2010:4;433-40.

27. Peay HL, Hollin I, Fischer R, Bridges JFP. A community-engaged approach to quantifying caregiver preferences for the benefits and risks of emerging therapies for Duchenne muscular dystrophy. Clin Ther. 2014;36(5):624-37.

28. Crowther CA, Dodd JM, Hiller JE, Haslam RR, Robinson JS, et al. Planned vaginal birth or elective repeat caesarean: patient preference restricted cohort with nested randomized trial. PLos Med. 2012;9(3): e1001192. doi: 10.1371/journal.pmed.1001192.

29. Long Q, Little RJ, Lin X. Causal inference in hybrid intervention trials involving treatment choice. J Am Stat Assoc. 2008;103:474-84.

30. Escudier B, Porta C, Bono P, Powles T, Eisen T, Sternberg CN, et al. Randomized, controlled, double-blind, cross-over trial assessing treatment preference for pazopanib versus sunitinib in patients with metastatic renal cell carcinoma: PISCES study. J Clin Oncol. 2014 May 10;32(14):1412-8.

31. Mitchell CC, Parikh OA. Factors involved in treatment preference in patients with renal cancer: pazopanib versus sunitinib. Patient Prefer Adherence. 2014;8:503-11.

32. Deal K. Segmenting patients and physicians using preferences from discrete choice experiments. Patient. 2014;7(1):5-21.

33. Liu SS, Chen J. Using data mining to segment healthcare markets from patients' preference perspectives. Int J Health Care Qual Assur. 2009;22(2):117-34.

34. O'Callaghan K, Shuren J. Listening to patients' views on new treatments for obesity. FDA Voice. 2015. Available at: http://blogs.fda.gov/fdavoice/index.php/tag/maestro-rechargeable-system/. Accessed March 29, 2015.

35. Ryan M, Watson V, Entwistle V. Rationalising the 'irrational': a think aloud study of discrete choice experiment responses. Health Econ. 2009;18:321-36.

36. Robinson S. Test-retest reliability of health state valuation techniques: the time trade off and person trade off. Health Econ. 2011;20(11):1379-91.

37. Salampessy BH, Veldwijk J, Jantine Schuit A, et al. The predictive value of discrete choice experiments in public health: an exploratory application. Patient. 2015 Jan 25. [Epub ahead of print]

38. Wood ME, Fama TA, Ashikaga T, Muss HB. Discrepancy between preference and actual adjuvant therapy for breast cancer. Clin Breast Cancer. 2010;10(5):398-403.

39. Hollin IL, Peay HL, Bridges JF. Caregiver preferences for emerging Duchenne muscular dystrophy treatments: a comparison of best-worst scaling and conjoint analysis. Patient. 2015;8(1):19-27.

40. Mühlbacher AC, Bethge S, Kaczynski A, Juhnke C. Patient's preferences regarding the treatment of type II diabetes mellitus: comparison of best-worst scaling and analytic hierarchy process. Value Health. 2013;16:A446.

41. Martin AJ, Glasziou PP, Simes RJ, Lumley T. A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. Int J Technol Assess Health Care. 2000;16(1):137-47.

42. Dolan JG. Multi-criteria clinical decision support: a primer on the use of multiple criteria decision making methods to promote evidence-based, patient-centered healthcare. Patient. 2010;3(4):229-48.

43. Felli JC, Noel RA, Cavazzoni PA. A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. Med Decis Making. 2009;29(1):104-15.

44. Dodgson JS, Spackman M, Pearman A, Phillips LD. Multi-criteria analysis: a manual. London: Department for Communities and Local Government; 2009.

45. Lynd LD, Marra CA, Najafzadeh M, Sadatsafavi M. A quantitative review of the regulatory assessment of the benefits and risks of rofecoxib relative to naproxen: an application of the incremental net-benefit framework. Pharmacoepidemiol Drug Saf. 2010;19(11):1172-80.

46. Mt-Isa S, Wang N, Hallgreen CE, Callreus T, Genov G, Hirsch I, et al.; PROTECT Work Package 5 participants. Review of methodologies for benefit and risk assessment of medication. Version 4. 2012. Available at: http://www.imi-protect.eu/documents/ShahruletalReviewofmethodologiesforbenefitandriskassessmentofmedicationMay2013.pdf. Accessed August 28, 2014.

47. Hughes D, Waddingham EAJ, Mt-Isa S, Goginsky A, Chan E, Downey G, et al.; IMI-PROTECT Work Package 5 participants. Recommendations for the methodology and visualisation techniques to be used in the assessment of benefit and risk of medicines. 2013. Available at: http://www.imi-protect.eu/documents/HughesetalRecommendationsforthemethodologyandvisualisationtechniquestobeusedintheassessmento.pdf. Accessed August 28, 2014.

48. Levitan B, Phillips LD, Walker S. Structured approaches to benefit-risk assessment: a case study and the patient perspective. Ther Innov Regul Sci. 2014;48(5):564-73.

49. Micaleff A, Callreus T, Phillips L, Hughes D, Hockley K, Wang N, et al.; PROTECT Work Package 5 participants. IMI work package 5: report 1:b:iii. Benefit - risk wave 1 case study report: Raptiva® (efalizumab). 2012. Available at: http://www.imi-protect.eu/documents/MicaleffAetalBenefitRiskWave1CasestudyReportEfalizumabFeb2013.pdf. Accessed August 22, 2014.

50. Quartey G, Hallgreen C, Chan E, Wang N, Lei G, Metcalf M; PROTECT Work Package 5 participants. IMI work package 5: report 1:b:ii. Benefit-risk wave 1 case study report: Ketek® (telithromycin). 2012. Available at: http://www.imi-protect.eu/documents/QuarteyetalBenefitRiskWave1CasestudyReportTelithromycinFeb2012.pdf. Accessed August 22, 2014.

51. Nixon R, Nguyen TST, Stoeckert I, Dierig C, Kuhls S, Hodgson G, et al.; PROTECT Work Package 5 participants. IMI WP5 report 1:b:iv. Benefit-risk wave 1 case study report: natalizumab. Risk benefit case study with a focus on testing methodology. 2013. Available at: www.imi-protect.eu/documents/NixonetalBenefitRiskWave1casestudyreportNatalizumabMay2013.pdf. Accessed August 22, 2014.

52. Juhaeri J, Mt-Isa S, Chan E, Genov G, Hirsch I, Bring J; PROTECT Work Package 5 participants. IMI work package 5: report 1:b:i. Benefit - risk wave 1 case study report:

rimonabant. 2011. Available at: http://www.imi-protect.eu/documents/JuhaerietalBenefitRiskWave1CasestudyreportRimonabantOct2011.pdf. Accessed August 22, 2014.

53. Phillips L, Amzal B, Asiimwe A, Chan E, Chen C, Hughes H, et al.; PROTECT Work Package 5 participants. IMI work package 5: report 2:b:ii. Benefit - risk wave 2 case study report: rosiglitazone. 2013. Available at: http://www.imi-protect.eu/documents/PhilipsetalBenefitRiskWave2CaseStudyReportRosiglitazoneFeb2013.pdf. Accessed August 22, 2014.

54. Nixon R, Waddingham E, Mt-Isa S, Hockley K, Elmachtoub A, Gelb D, et al.; PROTECT Work Package 5 participants. IMI PROTECT WP5 IMI report 2:b:iv. Natalizumab wave 2 case study report. Review of methodologies for benefit and risk assessment of medication. Risk benefit case study with a focus on testing methodology. 2013. Available at: www.imi-protect.eu/documents/NixonetalBenefitRiskWave2CasestudyReportNatalizumabMarch2013.pdf. Accessed August 22, 2014.

55. Johnson & Johnson Pharmaceutical Research & Development, LLC. Advisory committee briefing book: rivaroxaban for the prophylaxis of deep vein thrombosis (DVT) and pulmonary embolism (PE) in patients undergoing hip or knee replacement surgery. JNJ-39039039 (BAY 59-7939, rivaroxaban). February 12, 2009. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/UCM138385.pdf. Accessed August 28, 2014.

56. Torgerson DJ, Sibbald B. Understanding controlled trials. What is a patient preference trial? BMJ. 1998;316(7128):360.

57. Tilbrook H; Preference Collaborative Review Group. Patients' preferences within randomised trials: systematic review and patient level meta-analysis. BMJ. 2008;337:a1864. doi: 10.1136/bmj.a1864.

58. Marcus SM, Stuart EA, Wang P, Shadish WR, Steiner PM. Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. Psychol Methods. 2012;17(2):244-54.

59. Nixon R, Dierig C, Mt-Isa S, Stöckert I, Tong T, Kuhls S, et al. A case study using the PrOACT-URL and BRAT frameworks for structured benefit risk assessment. Biom J. 2015 Jan 26. doi: 10.1002/bimj.201300248. [Epub ahead of print]

60. Noel R, Herman R, Levitan B, Watson DJ, Van Goor K. Application of the Benefit-Risk Action Team (BRAT) Framework in Pharmaceutical R&D: results from a pilot program. Drug Inf J. 2012;46(6):736-43.

61. Levitan BS, Andrews EB, Gilsenan A, Ferguson J, Noel RA, Coplan PM, et al. Application of the BRAT Framework to Case Studies: Observations and Insights. Clin Pharmacol Ther. 2011;89(2):217-24.

62. Coplan PM, Noel RA, Levitan BS, Ferguson J, Mussen F. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. Clin Pharmacol Ther. 2011;89(2):312-5.

63. Mussen F, Salek S, Walker S, Phillips L. A quantitative approach to benefit-risk assessment of medicines - part 2: the practical application of a new model. Pharmacoepidemiol Drug Saf. 2007;16 Suppl 1:S16-41.