

Chapter 6. Outcome Definition and Measurement

Abstract

This chapter provides an overview of considerations for the development of outcome measures for observational comparative effectiveness research (CER) studies, describes implications of the proposed outcomes for study design, and enumerates issues of bias that may arise in incorporating the ascertainment of outcomes in observational research and means of evaluating, preventing and/or reducing these biases. Development of clear and objective outcome definitions that correspond to the nature of the hypothesized treatment effect and address the research questions of interest, along with validation of outcomes or use of standardized patient reported outcome (PRO) instruments validated for the population of interest, contribute to the internal validity of observational CER studies. Attention to collection of outcome data in an equivalent manner across treatment comparison groups is also required. Use of appropriate analytic methods suitable to the outcome measure and sensitivity analysis to address varying definitions of at least the primary study outcomes are needed to draw robust and reliable inferences. The chapter concludes with a checklist of guidance and key considerations for outcome determination and definitions for observational CER protocols.

Introduction

The selection of outcomes to include in observational comparative effectiveness research (CER) studies involves the consideration of multiple stakeholder viewpoints (provider, patient, payer, regulatory, industry, academic and societal) and the intended use of resulting evidence for decision making. It is also dependent on the level of funding and scope of the study. These studies may focus on clinical outcomes such as recurrence-free survival from cancer or coronary heart disease mortality, general health-related quality of life measures such as the EQ-5D and the SF-36 or disease-specific scales, like the uterine fibroid symptom and quality of life questionnaire (UFS-QOL), and/or health resource utilization or cost measures. As with other experimental and observational research studies, the hypotheses or study questions of interest must be translated to one or more specific outcomes with clear definitions.

The choice of outcomes to include in a CER study will in turn drive other important design considerations such as the data source(s) from which the required information can be obtained (see Chapter 8), the frequency and length of follow-up assessments to be included in the study following initial treatment, and the sample size, which is influenced by the expected frequency of the outcome in addition to the magnitude of relative treatment effects and scale of measurement.

In this chapter we provide an overview of types of outcomes with emphasis on those most relevant to observational CER studies, considerations in defining outcomes, the process of outcome ascertainment, measurement and validation, design and analysis considerations, and means to evaluate and address bias that may arise.

Conceptual models of health outcomes

In considering the range of health outcomes that may be of interest to patients, healthcare providers, and other decision-makers, medical conditions, impact to health-related or general quality of life, and resource utilization are key areas of focus. To address the interrelationships of these outcomes, some conceptual models have been put forth by researchers with a particular

focus on health outcomes studies. Two such models are described here.

Wilson and Cleary proposed a conceptual model or taxonomy integrating concepts of biomedical patient outcomes and measures of health related quality of life, which is divided into five levels: biological and physiological factors, symptoms, functioning, general health perceptions, and overall quality of life.¹ The authors discuss causal relationships between traditional clinical variables and measures of quality of life that address the complex interactions of biological and societal factors on health status, summarized in Table 6.1.

Table 6.1. Wilson and Cleary's Taxonomy of Biomedical and Health Related Quality of Life Outcomes

Level	Health concepts represented	Relationship with preceding level(s)
Biological and Physiological Factors	Genetic and molecular factors.	
Symptoms	Physical, psychosocial, emotional, and psychological symptoms.	Complex; symptoms may or may not be associated with biological or physiological factors (and vice versa).
Functional Status	Physical, social, role, psychological, and other domains of functioning.	Symptoms and biological and physiological factors are correlated with functional status, but may not completely explain variations. Other patient-specific factors (e.g., personality, social environment) are also important determinants.
General Health Perceptions	Subjective rating of general health.	Integrates all health concepts in the preceding levels; one of the best predictors of use of general medical and mental health services.
Overall Quality of Life	Summary measure of quality of life.	Although all preceding levels contribute to overall quality of life, general measures may not be strongly correlated with objective life circumstances as individuals may adjust expectations/goals with changing circumstances.

An alternative model, the ECHO (Economic, Clinical, Humanistic Outcomes) Model was developed for planning health outcomes and pharmaco-economic studies and goes a step further than the Wilson and Cleary model in incorporating costs and economic outcomes and their interrelationships with clinical and humanistic outcomes. The ECHO model does not explicitly incorporate characteristics of the patient as an individual or psychosocial factors to the extent of the Wilson and Cleary model however.

As suggested by the complex interrelationships between different levels and types of health outcomes, different terminology and classifications may be used and there are areas of overlap between the major categories of outcomes important to patients. In this chapter, we will discuss

outcomes according to the broad categories of Clinical, Humanistic, and Economic and Utilization outcome measures.

Outcome measurement properties

The properties of outcome measures that are an integral part of an investigator's evaluation and selection of appropriate measures include reliability, validity, and variability. **Reliability** is the degree to which a score or other measure remains unchanged upon test and re-test (when no change is expected), or across different interviewers or assessors. It is measured by statistics including kappa, and the inter- or intra-class correlation coefficient. **Validity**, broadly speaking, is the degree to which a measure assesses what it is intended to measure, and types of validity include face validity (the degree to which users or experts perceive that a measure is assessing what it is intended to measure), content validity (the extent to which a measure accurately and comprehensively measures what it is intended to measure), and construct validity (the degree to which an instrument accurately measures a non-physical attribute or construct such as depression or anxiety which is itself a means of summarizing or explaining different aspects of the entity being measured).² **Variability** usually refers to the distribution of values associated with an outcome measure in the population of interest, with a broader distribution or range of values said to show more variability.

Responsiveness is another property usually discussed in the context of PROs, but extendable to other measures, representing the ability of a measure to detect change in an individual over time.

All of these measurement properties may affect the degree of **measurement error** or **misclassification** that an outcome measure is subject to, with the consideration that the properties themselves are specific to the population and setting in which the measures are used. Issues of misclassification and considerations in reducing this type of error are discussed further in the section on "Avoidance of bias in study design".

Clinical outcomes

Clinical outcomes are perhaps the most common category of outcome to be considered in CER studies. Medical treatments are developed and must demonstrate efficacy in pre-approval clinical trials to prevent the occurrence of undesirable outcomes such as coronary events, osteoporosis, or death, to delay disease progression such as in rheumatoid arthritis, to hasten recovery or improve survival from disease, such as cancer or H5N1 influenza, or to manage or reduce the burden of chronic diseases including diabetes, psoriasis, Parkinson's, and depression. Post-approval observational CER studies are often needed to compare newer treatments against standard of care, to obtain real-world data on effectiveness as treatments are used in different medical care settings and broader patient populations than those studied in clinical trials, and to increase understanding of the relative benefits and risks of treatments by weighing quality of life, cost, and safety outcomes alongside clinical benefits. For observational studies, this category of outcomes generally focuses on clinically meaningful outcomes such as time between disease flares, number of swollen, inflamed joints, or myocardial infarction, though feasibility considerations sometimes dictate the use of intermediate.

Definitions of Clinical Outcomes

Temporal aspects

The nature of the disease state to be treated, the mechanism, and the intended effect of the treatment under study determine whether the clinical outcomes to be identified are incident (a first or new diagnosis of the condition of interest), prevalent (existing disease), or recurrent (new occurrence or exacerbation of disease in a patient who has a previous diagnosis of that condition). The disease of interest may be chronic (a long-term or permanent condition), acute (a condition with a clearly identifiable and rapid onset), transient (a condition that comes and goes) or episodic (a condition that comes and goes in episodes), or have more than one of these aspects.

Subjective vs. objective assessments

Most clinical outcomes involve a diagnosis or assessment by a health care provider. These may be recorded in a patient's medical record as part of routine care, coded as part of an electronic health record (EHR) or administrative billing system using systems such as ICD-9 or ICD-10, or collected specifically for a given study.

While there are varying degrees of subjectivity involved in most assessments by health care providers, objective measures are those that are not subject to a large degree of individual interpretation, and are likely to be reliably measured across patients in a study, by different health care providers, and over time. Laboratory tests may be considered objective measures in most cases and can be incorporated as part of a standard outcome definition to be used for a study when appropriate. Some clinical outcomes, such as all-cause mortality, can be ascertained directly and may be more reliable than measures that are subject to interpretation by individual health care providers, such as angina or depression.

Instruments have been developed to help standardize the assessment of some conditions for which a subjective clinical assessment might introduce unwanted variability. Consider the example of a study of a new psoriasis treatment. Psoriasis is a chronic skin condition that causes lesions affecting varying amounts of body surface area, with varying degrees of severity. While a physician may be able to assess improvement within an individual patient, a quantifiable measure that would be reproducible across patients and raters improves the information value of comparative trials and observational studies of psoriasis treatment effectiveness. An outcome assessment that relies on purely subjective assessments of improvement such as, "Has the patient's condition improved a lot, a little, or not at all?" is vulnerable to measurement error that arises from subjective judgments or disagreement among clinicians as to what comprises and how to rate the individual categories, often resulting in low reproducibility or inter-rater reliability of the measure. In the psoriasis example, an improved measure of the outcome would be a standardized assessment of the severity and extent of disease expressed as percentage of affected body surface area, such as the Psoriasis Area Severity Index or PASI Score.³ The PASI score requires rating the severity of target symptoms [erythema (E), infiltration (I), and desquamation (D)] and area of psoriatic involvement (A) for each of four many body areas [head (h), trunk (t), upper extremities (e), lower extremities (l)]. Target symptom severity is rated on a 0-4 scale; area of psoriatic involvement is rated on a 0-6 scale, with each numerical value representing a percentage of area involvement.³ The final calculated score ranges from 0 (no disease) to 72 (severe disease), with the score contribution of each body area weighted by its

percentage of total body area (10, 20, 30, and 40% of body area for head, upper extremities, trunk, and lower extremities, respectively).³ By using changes in the PASI score instead of subjective clinician assessments of overall performance, the PASI score increases reproducibility and comparability across studies that use the score.

Relatedly, the FDA has provided input on types of Clinical Outcome Assessments (COAs) that may be considered for qualification for use in clinical trials, with the goals of increasing the reliability of such assessments within a specific context of use in drug development and regulatory decision-making to measure a specific concept with a specific interpretation. Contextual considerations include the specific disease of interest, target population, clinical trial design and objectives, regionality, and mode of administration. The types of COAs described are:⁴

Patient-reported outcome (PRO) assessment: A measurement based on a report that comes directly from the patient (i.e., study subject) about the status of particular aspects of or events related to a patient's health condition. PROs are recorded without amendment or interpretation of the patient's response by a clinician or other observer. A PRO measurement can be recorded by the patient directly, or recorded by an interviewer provided that the interviewer records exactly the patient's response.

Observer reported outcome (ObsRO) assessment: An assessment that is determined by an observer who does not have a background of professional training that is relevant to the measurement being made, i.e., a non-clinician observer such as a teacher or caregiver. This type of assessment is often used when the patient is unable to self-report (e.g., infants, young children). An ObsRO assessment should only be used in the reporting of observable concepts (e.g., signs or behaviors); ObsROs cannot be validly used to directly assess symptoms (e.g., pain) or other unobservable concepts.

Clinician-reported outcome (ClinRO) assessment: An assessment that is determined by an observer with some recognized professional training that is relevant to the measurement being made.

Other considerations related to use of PROs for measurement of health-related quality of life and other concepts are addressed later on in this chapter.

Composite endpoints

Some clinical outcomes are composed of a series of items, and are referred to as composite endpoints. A composite endpoint is often used when the individual events included in the score are rare, and/or when it makes biological and clinical sense to group them. The study power for a given sample size may be increased when such composite measures are used as compared with individual outcomes, since by grouping numerous types of events into a larger category, the composite endpoint will occur more frequently than any of the individual components. As desirable as this can be from a statistical point of view, challenges include interpretation of composite outcomes that incorporate both safety and effectiveness and broader adoption of reproducible definitions that will enhance cross-study comparisons. For example, Kip et al.⁵ point out that there is no standard definition for MACE (major adverse cardiac events), a

commonly used outcome in clinical cardiology research. Kip and colleagues conducted analyses to demonstrate that varying definitions of composite endpoints, such as MACE, can lead to substantially different results and conclusions. The investigators utilized the DEScover registry patient population, a prospective observational registry of drug-eluting stent (DES) users, to evaluate differences in 1-year risk for three definitions of MACE in comparisons of patients with and without MI, and patients with multi-lesion stenting vs. single-lesion stenting (also referred to as percutaneous coronary intervention or PCI). The varying definitions of MACE included one related to safety only [composite of death, myocardial infarction (MI), and stent thrombosis (ST)], and two relating to both safety and effectiveness [composite of death, MI, ST, and either 1) target vessel revascularization (TVR) or 2) any repeat vascularization].⁵ When comparing patients with and without acute MI, the three definitions of MACE yielded very different hazard ratios. The safety only definition of MACE yielded a hazard ratio of 1.75 ($p < 0.05$), indicating that patients with acute MI were at greater risk of 1-year MACE. However, for the composite of safety and effectiveness endpoints, the risk of 1-year MACE was greatly attenuated and no longer statistically significant.⁵ Additionally, when comparing patients with single versus multiple lesions treated with PCI, the three definitions also yielded different results; while the safety only composite endpoint demonstrated that there was no difference in 1-year MACE, adding TVR to the composite endpoint definition led to a hazard ratio of 1.4 ($p < 0.05$) for multi-lesion PCI versus single-lesion PCI.⁵ This research serves as a cautionary tale for the creation and use of composite endpoints. Not only can varying definitions of composite endpoints such as MACE lead to substantially different results and conclusions, results must be carefully interpreted, especially in the case where safety and effectiveness endpoints are combined.

Intermediate endpoints

The use of an intermediate or surrogate endpoint is more common to clinical trials to observational studies. This type of endpoint is often a biological marker for the condition of interest, and may be used to reduce the follow-up period required to obtain results from a study of treatment effectiveness. An example would be the use of measures of serum lipids as endpoints in randomized trials of the effectiveness of statins, for which the major disease outcomes of interest to patients and physicians are a reduction in coronary heart disease incidence and mortality. The main advantages of intermediate endpoints are that the follow-up time required to observe possible effects of treatment on these outcomes may be substantially shorter than for the clinical outcome(s) of primary interest, and if they are measured on all patients, the number of outcomes for analysis may be larger. Similar to composite endpoints, using intermediate endpoints will increase study power for a given sample size as compared with outcomes that may be relatively rare, such as primary myocardial infarction. Surrogate or intermediate outcomes, however, may provide an incomplete picture of the benefits or risk. Treatment comparisons based on intermediate endpoints may differ in magnitude or direction from those based on major disease endpoints, as evidenced in a clinical trial of nifedipine versus placebo,^{6,7} as well as other clinical trials of antihypertensive therapy.⁹ On one hand, Nifedipine, a calcium channel blocker, was superior to placebo in reduction of onset of new coronary lesions; however, mortality was six-fold greater among patients who received Nifedipine versus placebo.⁶

Freedman and colleagues provided recommendations regarding the use of intermediate endpoints.⁸ Investigators should consider the degree to which the intermediate endpoint is

reflective of the main outcome, as well as the degree to which effects of the intervention may be mediated through the intermediate endpoint. Psaty and colleagues have cautioned that because drugs have multiple effects, to the extent that a surrogate endpoint is likely to measure only a subset of those effects, results of studies based on surrogate endpoints may be a misleading substitute for major disease outcomes as a basis for choosing one therapy over another.⁹

Table 6.2. Clinical Outcome Definitions and Objective Measures

Conceptual	Temporal aspects	Objective measure
Incident invasive breast cancer	Incident	SEER or state cancer registry data
Myocardial infarction	Acute, transient (in regard to elevated Troponin-I)	Review of laboratory test results for troponin and other cardiac enzymes for correspondence with a standard clinical definition
Psoriasis	Chronic, prevalent	Psoriasis Area Severity Index (PASI score) or percent body surface area assessment
Systemic lupus erythematosus (SLE)	Chronic condition with recurrent flares (episodes may have acute onset)	Systemic Lupus Erythematosus Disease Activity Index (SLEDAI)

Selection of clinical outcome measures

Identification of a suitable measure of a clinical outcome for an observational CER study is a process in which various aspects of the nature of the disease or condition under study should be considered along with sources of information by which the required information may be feasibly and reliably obtained.

The choice of outcome measure may follow directly from the expected biological mechanism of action of the intervention(s) under study and its impact on specific medical conditions. For example, the medications tamoxifen and raloxifene are selective estrogen receptor modulators that act through binding to estrogen receptors to block the proliferative effect of estrogen on mammary tissue and reduce the long-term risk of primary and recurrent invasive and non-invasive breast cancer.¹⁰ Broader or narrower outcome definitions may be appropriate to specific research questions or designs. In some situations, however, the putative biologic mechanism may not be well understood. Nonetheless, studies addressing the clinical question of comparative effectiveness of treatment alternatives may still inform decision making, and advances in understanding of the biological mechanism may follow discovery of an association through an observational CER study.

The selection of clinical outcome measures may be challenging when there are many clinical aspects that may be of interest, and a single measure or scale may not adequately capture the perspective of the clinician and patient. For example, in evaluating treatments or other interventions that may prolong the time between flares of SLE, researchers may use an index such as the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) which measures changes in disease activity. Or they may use the SLICC/ACR damage index, an instrument

designed to assess accumulated damage since the onset of the disease.^{11,12,13} This measure of disease activity has been tested in different populations and has demonstrated high reliability, evidence for validity, and responsiveness to change.¹⁴ Yet, multiple clinical outcomes may be of interest in studying treatment effectiveness in SLE, in addition to disease activity, such as reduction or increase in time to flare, reduction in corticosteroid use, or occurrence of serious acute manifestations (e.g., acute confusional state or acute transverse myelitis).¹⁵

Interactions with the health care system

One should first determine the source of reporting or detection for a medical condition that may lead to initial contact with the medical system. The manner in which the patient presents for medical attention may provide insights as to data source(s) that may be useful in studying the condition. The decision whether to collect information directly from the physician, through medical record abstraction (and where the relevant records might be found), directly from patients, and/or through use of electronic health records (EHRs) and/or administrative claims data will follow from this. For example, general hospital medical records are unlikely to provide the key components of an outcome such as respiratory failure which requires information about use of mechanical ventilation. In contrast, hospital medical records are useful for the study of myocardial infarction, which must be assessed and treated in a hospital setting and are nearly always accompanied by an overnight stay. General practice physician office records and emergency department records may be useful in studying the incidence of influenza A or urticaria, with selection of which of these depending on the severity of the condition. A prospective study may be required to collect clinical assessments of disease severity using a standard instrument, as these are not consistently recorded in medical practice and are not coded in administrative data sources. The Data Sources chapter (Chapter 8) provides additional information on selection of appropriate sources of data for an observational CE study.

Humanistic Outcomes

While outcomes of interest to patients generally include those of interest to physicians, payers, regulators and others, they are often differentiated by two characteristics: (1) they are clinically meaningful with practical implications for disease recognition and management (i.e., less interest in intermediate pathways with no clear clinical impact), and (2) they include reporting of outcomes based on a patient's unique perspective, e.g., patient reported scales that indicate the pain level, degree of functioning, etc. This section deals with measures of health-related quality of life (HRQoL) and the range of measures collectively described as patient-reported outcomes (PROs), which include measures of HRQoL. Other humanistic perspectives relevant to patients (e.g., economics, utilization of health services, etc.) are covered elsewhere.

Health related quality of life

Health-related quality of life (HRQOL) measures the impact of disease and treatment on the lives of patients and is defined as “the capacity to perform the usual daily activities for a person's age and major social role”.¹⁶ HRQOL commonly includes physical functioning, psychological well-being, and social role functioning. This construct comprises outcomes from the patient perspective and are measured by asking the patient or surrogate reporters about them.

HRQoL is an outcome increasingly used in randomized and non-randomized studies of health interventions, and as such the US Food and Drug Administration (FDA) has provided clarifying

definitions of HRQoL and of improvements in HRQoL. The FDA defines HRQoL as follows: “HRQL is a multi-domain concept that represents the patient’s general perception of the effect of illness and treatment on physical, psychological, and social aspects of life. Claiming a statistical and meaningful improvement in HRQL implies: (1) that all HRQL domains that are important to interpreting change in how the clinical trial’s population feels or functions as a result of the targeted disease and its treatment were measured; (2) that a general improvement was demonstrated; and (3) that no decrement was demonstrated in any domain.”¹⁷

Patient-reported outcomes

Patient-reported outcomes (PROs) include any outcome based on data provided by patients or people who can report on their behalf (proxies), as opposed to data provided from other sources.¹⁸ PROs refer to patient ratings and reports about any of several outcomes, including health status, health related quality of life, quality of life defined more broadly, symptoms, functioning, satisfaction with care, and satisfaction with treatment. Patients can also report about their health behaviors, including adherence and health habits. Patients may be asked to directly report information about clinical outcomes or health care utilization and out of pocket costs when these are difficult to measure through other sources. The FDA defines a PRO as “A measurement based on a report that comes directly from the patient (i.e., study subject) about the status of a patient’s health condition without amendment or interpretation of the patient’s response by a clinician or anyone else. A PRO can be measured by self-report or by interview provided that the interviewer records only the patient’s response.”¹⁷

In this section we focus mainly on the use of standard instruments for measurement of PROs, in domains including specific disease areas, health related quality of life, and functioning. PROs have similarities to other outcome variables measured in observational studies. They are measured with components of both random and systematic error (bias). To be most useful, it is important to have evidence about the reliability, validity, responsiveness, and interpretation of PRO measures, discussed further later in this section.

Types of humanistic outcome measures

Generic

Generic PRO questionnaires are designed to be used across different subgroups of individuals, and contain common domains that are relevant to almost all populations. They can be used to compare one population to another, or to compare scores in a specific population to normative scores. Many have been used for years, and have well established and well understood measurement properties.

Generic PRO questionnaires can focus on a comprehensive set of domains, or on a narrow range of domains, such as symptoms, or aspects of physical, mental, or social functioning. An example of a generic PRO is the Sickness Impact Profile (SIP), one of the oldest and most rigorously developed questionnaires, which measures 12 domains that are affected by illness.¹⁹ The SIP produces two subscale scores for Physical and Mental health, and an overall score. Another measure, the SF-36, measures 8 domains including general health perceptions, pain, physical functioning, role functioning (limited by physical health), social functioning, mental health, and vitality.²⁰ The SF-36 produces a Physical Component Score and a Mental Component Score.²¹ The EQ-5D is another generic measure of health-related quality of life, intended for self-

completion, that generates a single index score. This scale defines health in terms of 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has 3 response categories corresponding to no problem/some problem/extreme problem. Taken as a whole, the EQ-5D defines a total of 243 possible states to which 2 further states (dead and unconscious) were added.²² Another broadly used indicator of quality of life relates to the ability to work. The Work Productivity Index (WPAI) was created as a patient-reported quantitative assessment of the amount of absenteeism, presenteeism and daily activity impairment attributable to general health (WPAI:GH) or a specific health problem (WPAI:SHP) (see below), in an effort to develop a quantitative approach to measuring the ability to work.²³

Examples of generic measures that assess a more restricted set of domains include the SCL-90 to measure symptoms,²⁴ the Index of Activities of Daily Living to measure independence in performing basic functioning,²⁵ the Psychological General Well-Being Index to measure psychological well-being (PGWBI),²⁶ and the Beck Depression Inventory.²⁷

Disease or population specific

Specific PRO questionnaires are sometimes referred to as “Disease-Specific.” While a questionnaire can be disease- or condition-specific (e.g., chronic heart failure), it can also be designed for use in a specific population (e.g., pediatric, geriatric), or for use to evaluate a specific treatment (e.g., renal dialysis). Specific questionnaires may be more sensitive to symptoms that are experienced by a particular group of patients. Thus, they are thought to detect differences and changes in scores when they occur in response to interventions.

Some specific measures assess multiple domains that are affected by a condition. For example, the Arthritis Impact Measurement Scales (AIMS) includes 9 subscales that assess problems specific to the health-related quality of life of patients with rheumatoid arthritis and its treatments.²⁸ The MOS-HIV Health Survey includes 10 domains that are salient for people with HIV and its treatments.²⁹

Some of these measures take a modular approach, including a core measure that is used for assessment of a broader set of conditions, accompanied by modules that are specific to disease subtypes. For example, the FACIT and EORTC families of measures for evaluating cancer therapies each include a core module that is used for all cancer patients, and specific modules for each type of cancer, such as a module pertaining specifically to breast cancer.^{30,31,32}

Other measures focus more narrowly on a few domains most likely to be affected by a disease, or most likely to improve with treatment. For example, the Headache Impact Test includes only six items.³³ In contrast, other popular measures focus on symptoms that are affected by many diseases, such as the Brief Pain Inventory and the M. D. Anderson Symptom Inventory (MDASI), which measure the severity of pain and other symptoms, and the impact of symptoms on function, and have been developed, refined and validated in many languages and patient subgroups over three decades.^{34,35}

It is possible, though not always advisable, to design a new PRO instrument for use in a specific study. The process of developing and testing a new PRO measure can be lengthy – generally requiring at least a year in time – and there is no guarantee that a new measure will work as well

as more generic but better tested instruments. Nonetheless, it may be necessary in the case of an uncommon condition for which there are no existing PRO measures, for a specific cultural context that differs from the ones that have been studied before, and/or to capture effects of new treatments that may require a different approach to measurement. However, when possible, in these cases it is still prudent to include a PRO measure with evidence for reliability and validity, ideally in the target patient population, in case the newly designed instruments fail to work as intended. This approach will allow comparisons with the new measure to assess content validity if there is some overlap of the concepts being measured.

Item Response Theory (IRT) and Computer Adaptive Testing (CATS)

Item Response Theory is a framework for the development of tests and measurement tools, and assessing how well the tools work. CAT represents an area of innovation in measuring PROs. CAT allows items to be selected to be administered so that questions are relevant to the respondent and targeted to the specific level of the individual, with the last response determining the next question that is asked. Behind the scenes, items are selected from “item banks,” comprising collections of dozens to hundreds of questions that represent the universe of potential levels of the dimension of interest, along with an indication of the relative difficulty or dysfunction that they represent. For example, the PROMIS item bank for physical functioning includes 124 items that range in difficulty from getting out of bed to running several miles.³⁶ This individualized administration can both enhance measurement precision and reduce respondent burden.³⁷ Computer adaptive testing is based on IRT methods of scaling items and drawing subsets of items from a larger item bank.³⁸ Considerations around adaptive testing involve balancing the benefit of tailoring the set of items and measurements to the specific individual with the risk of inappropriate targeting or classification if items answered incorrectly early on determine the later set of items to which a subject is able to respond. PROMIS (Patient-Reported Outcomes Measurement Information System)³⁹ is a major NIH initiative that leverages these desirable properties for PROs in clinical research and practice applications.

Descriptive vs. Preference format

Descriptive questionnaires ask about general or common domains and complaints, and usually provide multiple scores. Preference based measures, generally referred to as utility measures, provide a single score, usually on a 0-1 scale, that represents the aggregate of multiple domains for an overall estimate of burden.

Most of the questionnaires familiar to clinical researchers fall into the category of descriptive measures, including all of those mentioned in the preceding paragraphs. Patients or other respondents are asked to indicate the extent to which descriptions of specific feelings, abilities or behaviors apply to them. Utility measures are discussed further in the following section.

Other attributes of PROs

Within each of the above options, there are several attributes of PRO instruments to consider. These include response format (numeric scales vs. verbal descriptors or visual analogue scales), the focus of what is being assessed (frequency, severity, impairment, all of the above), and recall period. Shorter, more recent recall periods more accurately capture the individual’s actual experience, but may not provide as good an estimate of their typical activities or experiences (not everyone vacuums or has a headache every day).

Content validity

Content validity is the extent to which a PRO instrument covers the breadth and depth of salient issues for the intended group of patients. If a PRO instrument is not valid with respect to its content, then there is an increased chance that it may fail to capture adequately the impact of an intervention. For example, in a study to compare the impact of different regimens for rheumatoid arthritis, a PRO that does not assess hand function could be judged to have poor content validity, and might fail to capture differences among therapies. The FDA addresses content validity as being of primary interest in assessing a PRO, with other measurement properties being secondary and defines content validity as follows: “Evidence from qualitative research demonstrating that the instrument measures the concept of interest including evidence that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use. Testing other measurement properties will not replace or rectify problems with content validity.”¹⁷

Content validity is generally assessed qualitatively rather than statistically. It is important to understand and consider the population being studied, including their usual activities and problems, the condition (especially its impact on the patient’s functioning), and the interventions being evaluated (including both their positive and adverse effects).

Responsiveness and minimally important difference

Responsiveness is a measure of a PRO instrument’s sensitivity to changes in health status or other outcome being measured. If a PRO is not sufficiently responsive, it may not provide adequate evidence of effectiveness in observational studies or clinical trials. Related to responsiveness is the minimally important difference that a PRO measure may detect. Both the patient’s and health care provider’s perspectives are needed to determine if the minimally difference detectable by an instrument is in fact of relevance to their overall health status.⁴⁰

Floor and ceiling effects

Poor content validity can also lead to a mismatch between the distribution of responses and the true distribution of the concept of interest in the population. For example, if questions in a PRO to assess ability to perform physical activities are too “easy” relative to the level of ability in the population, then the PRO will not reflect the true distribution. This problem can present as a “ceiling” effect, where a larger proportion of the sample reports no disability. Similarly, “floor” effects are seen when questions regarding a level of ability are skewed too difficult for the population and the responses reflect this lack of variability.

Interpretation

Clinicians and clinical researchers may be unfamiliar with how to interpret PRO scores. They may not understand or have reference to the usual distribution of scores of a particular PRO in a clinical or general population. Without knowledge of normal ranges, physicians may not know what cut-points of scoring indicate that action is warranted. Researchers will not know whether an observed difference between two groups is meaningful, and whether a given change within or between groups is important without reference values from a comparable population. The task of understanding the meaning of scores is made more difficult by the fact that different PRO measurement tools tend to use different scoring systems. For most questionnaires, higher scores imply better health, but for some, a higher score is worse. Some scales are scored from 0-1,

where 0 = dead, and 1=perfect health. Others are scores on a 0-100 scale where 0 is simply the lowest attainable score (i.e., the respondent indicates the “worst” health state in response to all of the questions) and 100 is the highest. Still others are “normalized” so that for example, a score of 50 represents the mean score for the healthy or non-diseased population, with a standard deviation of 10 points. It is therefore crucial for researchers and users of PRO data to understand the scoring system being used for an instrument and the expected distribution, including the distributional properties.

For some PRO instruments, particularly generic questionnaires that have been applied to large groups of patients over many years, there are population norms that have been collected and established. These can be used as a reference point. Scoring can also be recalculated and “normalized” to a “T-score” so that a specific score (often 50 or 100) corresponds to the mean score for the population, and a specific number of points (often 5 or 10) corresponds to 1 standard deviation unit in that population.

Selection of a PRO measure

There are a number of practical considerations to take into account when selecting PRO measures for use in a CER study. The measurement properties discussed in the preceding sections also require evaluation in all instances for the specific instrument selected, within a given population, setting and intended purpose.

Population

It is important to understand the target population that will be completing the PRO assessment. These may range from individuals who can self-report, to individuals requiring the assistance of a proxy or medical professional (children, mentally or cognitively limited, visually impaired). Some respondents may be ambulatory individuals living in the community, whereas others may be inpatients or institutionalized individuals.

If a PRO questionnaire is to be used in non-English speaking populations, or in multiple languages, it is necessary to have appropriate language/culturally adapted versions. One should have evidence for the reliability and validity of the translated/culturally adapted version, as applied to the concerned population. One should also have data showing the comparability of performance across different language and cultural groups. This is of special importance when pooling data across language versions, as in a multinational clinical trial or registry study.

Burden

It is important to match the respondent burden created by a PRO to the requirements of the population being studied. Patients with greater levels of illness or disability are less able to complete lengthy questionnaires. In some cases, the content or specific questions posed in a PRO may be upsetting or otherwise unacceptable to respondents. In other cases, a PRO may be too cognitively demanding, or written at a reading level that is above that for the intended population. The total burden of study related data collection on patients and providers must also be considered, as an excessive number of forms that must be completed are likely to reduce compliance.

Cost and copyright

Another practical consideration is the copyright status of a PRO being considered for use. Some PROs are entirely in the public domain and are free for use. Others are copyrighted and require permission and/or the payment of fees for use. Some scales require payment of fees for scoring, such as the SF-12 and SF-36.

Mode and format of administration

As noted above, there are various options for how a questionnaire should be administered, and how the data should be captured, each method having both advantages and disadvantages. A PRO can be 1) self-administered at the time of a clinical encounter, 2) administered by an interviewer at the time of a clinical encounter, 3) administered with computer assistance at the time of a clinical encounter, 4) self-administered by mail, 5) self-administered on-line, 6) interviewer administered by phone, and 7) computer administered by phone. Self-administration at the time of a clinical encounter requires little technology or up-front cost, but requires staff for supervision and data entry and can be difficult for respondents with limited literacy or sophistication. Face to face administration engages respondents and reduces their burden, but requires trained interviewers. Computer assisted provides an intermediate solution but also requires capital investment. Mailed surveys afford more privacy to respondents, but generate expenses related to mailing, and do not eliminate problems with literacy. Paper-based formats require data entry, scoring and archiving and is prone to calculation errors. Online administration is relatively inexpensive, especially for large surveys, and surveys can be completed any time, but not all individuals have internet access. Live telephone interview is engaging and allows interviewer flexibility, but is also expensive. “Cold calls” to potential study participants may result in low response rates given the rise of caller ID systems to screen calls, and skepticism about “telemarketing”.

Interactive voice response systems (or IVRS) can also be used to conduct telephone interviews, but it can be tedious to respond using the phone key pad and this format strikes some as impersonal.

Static versus dynamic questionnaires

Static forms are the type of questionnaire that employs a fixed format set of questions and response options. They can be administered on paper, by interview, or through the internet. Dynamic questionnaires select follow-up questions to administer based on the responses already obtained for previous question, and since they are more efficient, more domains can be assessed.

Economic and Utilization Outcomes

While clinical outcomes represent the provider professional perspective and humanistic outcomes represent the patient perspective, economic outcomes, including measures of health resource utilization represent the payer and societal perspective. Measures of cost and cost-effectiveness are often excluded from government-funded CER studies in the US. However, these measures are important to a variety of important stakeholders such as payers and product manufacturers, and are routinely included in cost effectiveness research in countries such as Australia, the United Kingdom, Canada, France, and Germany.⁴¹

Research questions addressing issues of cost-effectiveness and resource utilization may be formulated in a number of ways. Cost identification studies measure the cost of applying a specified treatment to a population under a certain set of conditions. These studies describe the cost incurred without comparison to alternative interventions. Some cost identification studies describe the total costs of care for a particular population whereas others isolate costs of care related to the specific condition; this latter approach requires that each episode of care be ascribed as having been related or unrelated to the illness of interest and involves substantial review.⁴² Cost benefit studies are typically measured in dollars or other currency. These studies compare the monetary costs of an intervention against the standard of care with the cost savings that result from the benefits of that treatment. In these studies, mortality is also assigned a dollar value, although techniques for assigning value to a human life are controversial. Cost effectiveness is a relative concept and its analysis compares the costs of treatments and benefits of treatments in terms of a specified outcome, such as reduced mortality or morbidity, such as years of life saved, or infections averted.

Types of health resource utilization and cost measures

Monetary costs

Studies most often examine direct costs (the monetary costs of medical treatments themselves, potentially including associated costs of administering treatment or conditions associated with treatment) but may also include measures of indirect costs (the costs of disability or loss of livelihood, both actual and potential). Multiple measures of costs are commonly included in any given study.

Health resource utilization

Measures of health resource utilization, such as number of inpatient or outpatient visits, total days of hospitalization in a given year, or number of days treated with IV antibiotics are often used as efficient and easily interpretable proxies for measuring cost, since actual costs are dependent on numerous factors (e.g., institutional overhead, volume discounts) and can be difficult to obtain since they often may be confidential since, in part, they reflect business acumen in price negotiation. Costs may also vary by institution or location, such as the cost of a day in the hospital or a medical procedure. Resource utilization measures may be preferred when a study is intended to yield results that may be generalizable to other health systems or reimbursement systems than those under study, as they are not dependent on a particular reimbursement structure such as Medicare. Alternatively, a specific cost or reimbursement structure, such as the amount reimbursed by the Center for Medicare and Medicaid Services (CMS) for specific treatment items or average wholesale drug costs, may be applied to units of health resource use when conducting studies that pool data from different health systems.

Utility and preference-based measures

PROs and cost analyses intersect around the calculation of cost-utility. Utility measures are derived from economic and decision theory. The term utility refers to the value placed by the individual on a particular health state. Utility is summarized as a score ranging from 0.0 representing death to 1.0 representing perfect health.

In health economic analyses, utilities are used to justify devoting resources to a treatment. There

are several widely used preference based instruments that are used to estimate utility.

Preference measures are based on the fundamental concept that individuals or groups have reliable preferences about different health states. To evaluate those preferences, individuals rate a series of health states: for example, a person with specific levels of physical functioning (able to walk one block but not climb stairs), mental health (happy most of the time), and social role functioning (not able to work due to health). The task for the individual is to directly assign a degree of preference to that state. These include the Standard Gamble and Time Tradeoff methods,^{43,44} the EQ-5D, also referred to as the Euroqol,²² the Health Utilities Index,^{45,46} and the Quality of Well-being Scale.⁴⁷

Quality-adjusted life years (QALYs)

Utility scores associated with treatment can be used to weight the duration of life according to its quality, and are then used to generate QALYs. Utility scores are generally first ascertained directly in a sample of people with the condition in question, either cross-sectionally or over time with a clinical trial. Utility values are sometimes estimated indirectly using other sources of information about the health status of people in a population. The output produced by an intervention can be calculated as the area under the cost-utility curve.

For example, if the mean utility score for patients receiving antiretroviral treatment for HIV disease is 0.80, then the outcome for a treated group would be survival time multiplied by 0.80.

Disability-adjusted life years (DALYs)

DALYs are another measure of overall disease burden expressed as the number of years lost to poor health, disability, or premature death.⁴⁸ As with QALYs, mortality and morbidity are combined in a single metric. Potential years of life lost to premature death are supplemented with years of health life lost due to less than optimal health. Whereas 1 QALY corresponds to one year of life in optimal health, 1 DALY corresponds to one year of healthy life lost.

An important aspect of the calculation of DALYs is that the value assigned to each year of life depends on age. Years lived as a young adult are valued more highly than those spent as a young child or older adult, reflecting the capacity for work productivity during different phases of life. DALYs are therefore estimated for different chronic illnesses by first calculating the age and sex adjusted incidence of disease. A DALY is calculated as the sum of the average years of life lost, and the average years lived with a disability. For example, to estimate the years of healthy life lost in a region due to HIV/AIDS, one would first estimate the prevalence of the disease by age. The DALY value is calculated by summing the average of years of life lost and the average number of years lived with AIDS, discounted based on a universal set of standard weights based on expert valuations.

Selection of resource utilization and cost measures

The selection of measures of resource utilization or costs should correspond to the primary hypothesis in terms of the impact of intervention. For example, will treatment reduce the need for hospitalization or result in a shorter length of stay? Or, will treatment or intervention reduce complications that require hospitalization? Or, will a screening method reduce the total number of diagnostic procedures required per diagnosis?

It is useful to consider what types of costs are of interest to the investigators and to various stakeholders. Are total costs of interest, or costs associated with specific resources (e.g., prescription drug costs)? Are only direct costs being measured, or are you also interested in indirect costs such as those related to days lost from work?

When it is determined to present results in terms of dollars rather than units of resources, several different methods can be applied. In the unusual case that an institution has a cost-accounting system, cost can be measured directly. In most cases, resource units are collected, and assigned costs based on local or national average prices for the specific resources being considered, e.g., reimbursement from CMS for a CT scan, or a hospital day. Application of an external standard cost system reduces variability in costs due to region, payer source, and other variables that might obscure the impact of the intervention in question.

Study Design and Analysis Considerations

Study period and length of follow-up

In designing a study, the required study period and length of follow-up are determined by the expected timeframe within which an intervention may be expected to impact the outcome of interest. A study comparing traditional with minimally invasive knee replacement surgery will need to follow subjects at least for the duration of the expected recovery time of three to six months or longer. The optimal duration of a study can be problematic when studying effects that may manifest over a long time period, such as treatments to delay onset or prevent chronic disease. In these cases, data sources with a high degree of turnover in patients, such as administrative claims data bases from managed care organizations, may not be suitable. For example, in the case of Alzheimer's disease, a record of health care is likely to be present in health insurance claims. However, with the decline in cognitive function, patients may lose ability to work and enter assisted care facilities, where utilization is not typically captured in large health insurance claims systems. Some studies may be undertaken for the purpose of determining how long an intervention can be expected to impact the outcome of interest. For example, various measures are used to aid in reducing obesity and smoking cessation, and patients, health care providers, and payers are interested in knowing how long these interventions work (if at all), for whom, and in what situations.

Notwithstanding the limitations of intermediate endpoints (discussed in a preceding section), one of the main advantages of their use is the potential truncation of the required study follow-up period. Consider, for example, a study of the efficacy of the human papilloma virus vaccine, for which the major medical endpoint of interest is prevention of cervical cancer. The long latency period (more than two years, depending on the study population) and relative infrequency of cervical cancer raise the possibility that intermediate endpoints should be used. Candidates might include new diagnoses of genital warts, or new diagnoses of the precancerous conditions cervical intraepithelial neoplasia (CIN) or vaginal intraepithelial neoplasia (VIN), which have shorter latency periods of less than one year or two years (minimum) respectively. Use of these endpoints would allow such a study to provide meaningful evidence informing the use of the HPV vaccine in a shorter timeframe, during which more patients might benefit from its use. Alternatively, if the vaccine is shown to be ineffective, this could avoid years of unnecessary

treatment and the associated costs as well as the costs of running a longer trial.

Avoidance of bias in study design

Misclassification

The role of the researcher is to understand the extent and sources of misclassification in outcome measurement, and to try to reduce these as much as possible. To ensure comparability between treatment groups with as little misclassification (also referred to as measurement error) of outcomes as possible, a clear and objective (verifiable and not subject to individual interpretation insofar as possible) definition of the outcome of interest is needed. An unclear outcome definition can lead to misclassification and bias in the measure of treatment effectiveness. When the misclassification is non-differential, or equivalent across treatment groups, the estimate of treatment effectiveness will be biased toward the null, reducing the apparent effectiveness of treatment, which may result in an erroneous conclusion that no effect (or one smaller than the true effect size) exists. When the misclassification differs systematically between treatment groups, it may distort the estimate of treatment effectiveness in either direction.

For clinical outcomes, incorporation of an objective measure such as a validated tool that has been developed for use in clinical practice settings, or an adjudication panel for review of outcomes with regard to whether they meet the pre-determined definition of an event, would both be approaches that increase the likelihood that outcomes will be measured and classified accurately and in a manner unlikely to vary according to who is doing the assessment. For PROs, measurement error can stem from several sources, including the way in which a question is worded and hence understood by a respondent, how the question is presented, the population being assessed, the literacy level of respondents, the language in which the questions are written, and elements of culture that it represents.

To avoid differential misclassification of outcomes, care must also be taken to use the same methods of ascertainment and definitions of study outcomes whenever possible. For prospective or retrospective studies with contemporaneous comparators, this is usually not an issue since it is most straightforward to utilize the same data sources and methods of outcome ascertainment for each comparison group. A threat to validity may arise in use of a historical comparison group, which may be used in certain circumstances. For example, this occurs when a new treatment largely displaces use of an older treatment within a given indication, but further evidence is needed for the comparative effectiveness of the newer and older treatments, such as enzyme replacement for lysosomal storage disorders. In such instances, use of the same or similar data sources, and equivalent outcome definitions to the extent possible will reduce the likelihood of bias due to differential outcome ascertainment.

Other situations that may give rise to issues of differential misclassification of outcomes include: when investigators are not blinded to the hypothesis of the study, and “rule-out” diagnoses are more common in those with a particular exposure of interest; when screening or detection of outcomes is more common or more aggressive in those with one treatment than another (i.e., surveillance bias, e.g., when liver function testing are preferentially performed in patients using a new drug compared to other treatments for that condition); and when loss to follow-up occurs that is related to the risk of experiencing the outcome. For example, once a safety signal has

been identified and publicized, physicians have been alerted and then look more proactively for clinical signs and symptoms in treated patients. This situation is even greater for products that are subject to controlled distribution or Risk Evaluation and Mitigation Strategies (REMS). Consider clozapine, an anti-schizophrenia drug that is subject to controlled distribution through a “no blood, no drug” monitoring program. The blood testing program was implemented to detect early development of agranulocytemia. When comparing patients treated with clozapine to other anti-schizophrenics, those using clozapine may appear to have a worse safety profile with respect to this outcome.

Validation and adjudication

In some instances, additional information must be collected (usually from medical records) to validate the occurrence of the outcome of interest, including to exclude erroneous or “rule-out” diagnoses. This is particularly important for medical events identified in administrative claims databases, for which a diagnosis code associated with a medical encounter may represent a “rule out” diagnosis or a condition that does not map to a specific diagnosis code. For some complex diagnoses, such as unstable angina, a standard clinical definition must be applied by an adjudication panel that has access to detailed records inclusive of subjects’ relevant medical history, symptomatic presentation, diagnostic work-up, and treatment. Methods of validation and adjudication of outcomes strengthen the internal validity and therefore the evidence that can be drawn from a CER study. However, they are resource intensive.

Issues specific to PROs

PROs are prone to several specific sources of bias. Self-reports of health status are likely to differ systematically from reports by surrogates, who, for example, are likely to report less pain than the individuals themselves.⁴⁹ Some biases may be population dependent. For example, there may be a greater tendency of some populations to succumb to acquiescence bias (agreeing with the statements in a questionnaire) or social desirability bias (answering in a way that would cast the respondent in the best light).⁵⁰ In some situations, however, PRO may be the most useful marker of disease activity, such as with episodic conditions that cause short duration disease flares such as low back pain and gout, where patients may not present for health care immediately, if at all.

The goal of the researcher is to understand and reduce sources of bias, considering those most likely to apply in the specific population and topics under study. In the case of well-understood systematic biases, adjustments can be made so that distributions of responses are more consistent. In other cases, redesigning items and scales, for example, including both positively and negatively worded items, can reduce specific kinds of bias.

Missing data, an issue covered in more detail in Chapter 10, pose a particular problem with PROs, since PRO data are usually not missing at random. Instead, respondents whose health is poorer are more likely to fail to complete an assessment. Another special case of missing data occurs when a patient dies and is unable to complete an assessment. If this issue is not taken into account in the data analysis, and scores are only recorded for living patients, incorrect conclusions may be drawn. Strategies for handling this type of missing data include selection of an instrument that incorporates a score for death, such as the Sickness Impact Profile,^{19,51} or the Quality of Well Being Scale,⁴⁷ or through an analytic strategy that allows for some missing

values.

Failure to account for missing PRO data that are related to poor health or death will lead to an overestimate of the health of the population based on responses from subjects who do complete PRO forms. Therefore in research using PROs, it is very important to understand the extent and pattern of missing data, both at the level of the individual as well as for specific items or scales on an instrument.

A strategy should be put in place to handle missing data when developing the study protocol and analysis plans. Such strategies that pertain to use of PROs in research are discussed in further detail in publications such as the book by Fairclough and colleagues.⁵²

Analytic considerations

Form of outcome measure and analysis approach

To a large extent, the form of the primary outcome of interest, that is, whether the outcome is measured and expressed as a dichotomous or polytomous categorical variable, a continuous variable, and whether it is to be measured at a single time point, repeated measures at fixed intervals, or repeated measures at varying time intervals, determines the appropriate statistical methods that may be applied in analysis. These topics are covered in detail in Chapter 10.

Sensitivity analysis

One of the key factors to address in planned sensitivity analyses for an observational CER study is how varying definitions of the study outcome or related outcomes will affect the measures of association from the study. These include assessing multiple related outcomes within a disease area, for example, multiple measures of respiratory function such as FEV1, FEV1% predicted, and FVC in studies of asthma treatment effectiveness in children, assessing the effect of different cutoffs for dichotomized continuous outcome measures such as use of Systemic Lupus Erythematosus Disease Activity Index-2000 scores to define active disease in lupus treatment studies,⁵³ or different sets of diagnosis codes to capture a condition in administrative data, such as influenza and related respiratory conditions. These and other considerations for sensitivity analyses are covered in detail in Chapter 11.

Conclusion

Future Directions

Increased use of EHRs as a source of data for observational research including registries, other types of observational studies, and specifically for CER has prompted initiatives to develop standardized definitions of key outcomes and other data elements that would be used across health systems and different EHR platforms to facilitate comparisons between studies and pooling of data. The National Cardiovascular Research Infrastructure partnership between the American College of Cardiology and Duke Clinical Research Institute that received ARRA funding to establish intra-operable data standards based on the National Cardiovascular Data Registry is an example of such a current activity.⁵⁴

Summary

This chapter provides an overview of considerations in development of outcome definitions for observational CER studies, describes implications of the nature of the proposed outcomes for the study design, and enumerates issues of bias that may arise in incorporating the ascertainment of outcomes in observational research and means of preventing or reducing these biases.

Development of clear and objective outcome definitions that correspond to the nature of the hypothesized treatment effect and address the research questions of interest, along with validation of outcomes where warranted or use of standardized PRO instruments validated for the population of interest, contribute to the internal validity of observational CER studies. Attention to collection of outcome data in an equivalent manner across treatment comparison groups is also required. Use of appropriate analytic methods suitable to the outcome measure and sensitivity analysis to address varying definitions of at least the primary study outcomes are needed to make inferences drawn from such studies more robust and reliable.

DRAFT

Checklist: Guidance and Key Considerations for Outcome Selection and Measurement for Observational CER Protocols and Proposals

Guidance	Key Considerations	Check
Propose primary and secondary outcomes that directly correspond to research questions	<ul style="list-style-type: none"> - Follow-up period should be sufficient to observe hypothesized effects of treatment on primary and secondary outcomes 	<input type="checkbox"/>
Provide clear and objective definitions of clinical outcomes	<ul style="list-style-type: none"> - Should reflect hypothesized mechanism of effect of treatment, if known - Should provide justification that outcome is reliably ascertained without additional validation, when applicable and feasible, or propose validation and/or adjudication of endpoints. - If an intermediate (surrogate) endpoint is proposed, justification should be provided why main disease outcome of interest is not being used and that intermediate endpoint reflects expected pathway of effect of treatment on main outcome of interest 	<input type="checkbox"/>
Provide clear and relevant definitions of cost or health resource utilization outcomes	<ul style="list-style-type: none"> - Should reflect hypothesized effect of treatment on specific components of medical cost and/or resource utilization, if known - Should be able to be measured directly or via proxy from data sources proposed for study - For costs, should consider proposing standard benchmark costs to be applied to units of resource utilization especially when multiple health systems, payment systems, and/or geographic regions are included in study population or data source 	<input type="checkbox"/>
Describe plan for use of validated, standard instrument for measurement of patient-reported-outcomes	<ul style="list-style-type: none"> - Should reflect hypothesized effect of treatment on specific aspects of disease symptoms or treatment, or quality of life, if known - Should propose use of a standard instrument that has been validated for use in population representative of the study population, when possible - Should be validated for use in translation to other specific languages if intended to be used in those languages for study, when possible - Should be validated for the intended mode of administration, when possible 	<input type="checkbox"/>
Address issues of bias expected to arise and proposed means of bias minimization	<ul style="list-style-type: none"> - Describe potential issues of bias, misclassification, and missing data that may be expected to occur with the proposed outcomes, including those specific to PRO data - Provide plan for minimization of potential bias, misclassification, and missing data issues identified 	<input type="checkbox"/>
Analysis	<ul style="list-style-type: none"> - Proposed analytic methods should correspond to nature of outcome measure (e.g., continuous, categorical [dichotomous, polychotomous, ordinal], repeated measures, time-to-event) - Sensitivity analyses relating to expected questions that arise around the study outcomes 	<input type="checkbox"/>

Guidance	Key Considerations	Check
	<ul style="list-style-type: none">- Sensitivity analyses should be proposed that address different relevant definitions of the study outcome(s) or multiple related outcomes (e.g., different measures of subclinical and clinical cardiovascular disease)	

DRAFT

References

- ¹ Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995 Jan 4; 273:59-65.
- ² Streiner DL, Norman GR. *Health Measurement Scales: a Practical Guide to their Development and Use*. 4th ed. Oxford University Press; 2008.
- ³ Fredriksson T, Pettersson U. Severe psoriasis--oral therapy with a new retinoid. *Dermatologica* 1978;157(4):238-44.
- ⁴ US Department of Health and Human Services, Food and Drug Administration. Clinical Outcome Assessment Qualification Program. April 2012. Available at: [http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm?utm_source=fdaSearch&utm_medium=website&utm_term=drug development tools qualification program&utm_content=2](http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm?utm_source=fdaSearch&utm_medium=website&utm_term=drug%20development%20tools%20qualification%20program&utm_content=2). Accessed April 16, 2012.
- ⁵ Kip KE, Hollabaugh K, Marroquin OC, Williams DO. The problem with composite end points in cardiovascular studies. The story of major adverse cardiac events and percutaneous coronary intervention. *J Am Coll Cardiol* 2008;51:701-7.
- ⁶ Lichtlen PR, Hugenholz PG, Rafflenbeul W, et al. Retardation of angiographic progression of coronary artery disease by nifedipine: results of the International Nifedipine Trial on Antiatherosclerotic Therapy (INTACT). *Lancet* 1990;335:1109-1113.
- ⁷ Psaty BM, Siscovick DS, Weiss NS, et al. Hypertension and outcomes research. From clinical trials to clinical epidemiology. *Am J Hypertens* 1996;9:178-183.
- ⁸ Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992 Jan 30;11(2):167-78.
- ⁹ Psaty BM, Lumly T. Surrogate end points and FDA approval: A tale of 2 lipid-altering drugs. *JAMA* 2008; 299(12):1474-6.
- ¹⁰ Vogel VG, Costantino JP, Wickerham DL, et al. Update of the National Surgical Adjuvant Breast and Bowel Project Study of Tamoxifen and Raloxifene (STAR) P-2 Trial: Preventing breast cancer. *Cancer Prev Res (Phila)* 2010 Jun;3(6):696-706.
- ¹¹ Gladman DD and Urowitz MB. "The SLICC/ACR damage index: progress report and experience in the field." *Lupus* 1999;8:632-637.
- ¹² Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang DH, and the Committee on Prognosis Studies in SLE. Derivation of the SLEDAI: a disease activity index for lupus patients. *Arthritis Rheum* 1992;35:630-40.
- ¹³ Gladman DD, Ibanez D, Urowitz MB. Systemic Lupus Erythematosus Disease Activity Index 2000. *J Rheumatol* 2002;29:288-91.
- ¹⁴ Griffiths B, Mosca M, Gordon C. Assessment of patients with systemic lupus erythematosus and the use of lupus disease activity indices. *Best Pract Res Clin Rheumatol* 2005 Oct;19(5):685-708.
- ¹⁵ US Department of Health and Human Services, Food and Drug Administration. Guidance for Industry Systemic Lupus Erythematosus — Developing Medical Products for Treatment. June 2010. Available at: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072063.pdf>. Accessed February 3, 2012.
- ¹⁶ Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993 Apr 15;118(8):622-9.
- ¹⁷ U.S. Department of Health and Human Services. FDA. Guidance for Industry Patient Reported Outcome Measures: Use in medical product development to support labelling claims. December 2009.

- ¹⁸ Acquadro C, Berzon R, Dubois D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. *Value Health* 2003 Sep;6:522-31.
- ¹⁹ Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981 Aug;19(8):787-805.
- ²⁰ Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992 Jun;30(6):473-83.
- ²¹ Ware JE Jr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Med Care* 1995 Apr;33(4 Suppl):AS264-79.
- ²² EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. *Health Policy* 1990 Dec;16(3):199-208.
- ²³ Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics* 1993 Nov;4(5):353-65.
- ²⁴ Derogatis LR, Cleary PA. Factorial invariance across gender for the primary symptom dimensions of the SCL-90. *Br J Soc Clin Psychol* 1977 Nov;16(4):347-56.
- ²⁵ Katz S, Akpom CA. 12. Index of ADL. *Med Care* 1976 May;14(5 Suppl):116-8.
- ²⁶ Dupuy HJ. The Psychological general Well-Being (PGWB) Index. In: *Assessment of Quality of Life in clinical trials of cardiovascular therapies*. Edited by Wenger NK, Mattson ME, Furberg CD, Elinson J. Le Jacq Publishing 1984; Chap 9:170-183
- ²⁷ Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. *Arch Gen Psychiatry* 1961;4:561-71.
- ²⁸ Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The arthritis impact measurement scales. *Arthritis Rheum* 1980 Feb;23(2):146-52.
- ²⁹ Wu AW, Revicki DA, Jacobson D, Malitz FE. Evidence for reliability, validity and usefulness of the Medical Outcomes Study HIV Health Survey (MOS-HIV). *Qual Life Res* 1997 Aug;6(6):481-93.
- ³⁰ Cella D, Nowinski CJ. Measuring quality of life in chronic illness: the functional assessment of chronic illness therapy measurement system. *Arch Phys Med Rehabil* 2002 Dec;83(12 Suppl 2):S10-7.
- ³¹ Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993 Mar 3;85(5):365-76.
- ³² Sprangers MA, Cull A, Groenvold M, Bjordal K, Blazeby J, Aaronson NK. The European Organization for Research and Treatment of Cancer approach to developing questionnaire modules: an update and overview. EORTC Quality of Life Study Group. *Qual Life Res* 1998 May;7(4):291-300.
- ³³ Kosinski M, Bayliss MS, Bjorner JB, et al. A six-item short-form survey for measuring headache impact: the HIT-6. *Qual Life Res*. 2003 Dec;12(8):963-74.
- ³⁴ Cleeland CS. Symptom burden: multiple symptoms and their impact as patient-reported outcomes. *J Natl Cancer Inst Monogr* 2007;37:16-21.
- ³⁵ Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. *Ann Acad Med Singapore* 1994;23(2):129-138.
- ³⁶ Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res* 2011;29(6):947-53.
- ³⁷ Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. *Qual Life Res* 2007;16 Suppl 1:95-108.

- ³⁸ Reise SP. Item response theory: fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*. 2005 April;14(2):95-101.
- ³⁹ Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 2007 May;45:S22-S31.
- ⁴⁰ Revicki DA, Cella D, Hays RD, et al. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006; 4:70.
- ⁴¹ Chalkidou K, Tunis S, Lopert R, et al. Comparative effectiveness research and evidence-based health policy: experience from four countries. *Milbank Q* 2009 Jun;87(2):339-67.
- ⁴² Lanes SF, Lanza LL, Radensky, et al. Resource utilization and cost of care for rheumatoid arthritis and osteoarthritis in a managed care setting: the importance of drug and surgery costs. *Arthritis and Rheumatism* 1997;40(8):1475-1481.
- ⁴³ Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ* 1986 Mar;5(1):1-30.
- ⁴⁴ Torrance GW. Utility approach to measuring health-related quality of life. *J Chronic Dis* 1987;40(6):593-603.
- ⁴⁵ Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. *Pharmacoeconomics* 1995 Jun;7(6):490-502.
- ⁴⁶ Feeny D, Furlong W, Saigal S, Sun J. Comparing directly measured standard gamble scores to HUI2 and HUI3 utility scores: group- and individual-level comparisons. *Soc Sci Med* 2004 Feb;58(4):799-809.
- ⁴⁷ Kaplan RM, Anderson JP. The General Health Policy Model: an integrated approach. In: Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. *Med Care* 2000 Sep;38:II138-II150.
- ⁴⁸ Murray CJ. Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bull World Health Organ* 1994;72(3):429-45.
- ⁴⁹ Wilson KA, Dowling AJ, Abdolell M, Tannock IF. Perception of quality of life by patients, partners and treating physicians. *Qual Life Res* 2000;9(9):1041-52.
- ⁵⁰ Ross CK, Steward CA, Sinacore JM. A comparative study of seven measures of patient satisfaction. *Med Care* 1995 Apr;33(4):392-406.
- ⁵¹ Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS. The sickness impact profile: validation of a health status measure. *Med Care* 1976 Jan;14:57-67.
- ⁵² Fairclough DL. *Design and Analysis of Quality of Life Studies in Clinical Trials*. 2nd ed. Boca Raton: Chapman and Hall/CRC Press; 2010.
- ⁵³ Yee CS, Farewell VT, Isenberg DA, et al. The use of Systemic Lupus Erythematosus Disease Activity Index-2000 to define active disease and minimal clinically meaningful change based on data from a large cohort of systemic lupus erythematosus patients. *Rheumatology (Oxford)*. 2011 May;50(5):982-8.
- ⁵⁴ National Cardiovascular Research Infrastructure (NCRI). Available at: <https://www.ncrinetwork.org/>. Accessed February 3, 2012.