Considerations and Challenges in Selecting Patient-Reported Outcome Measures for Clinical Trials in Nephrology

Angela Ju and Allison Tong


Introduction

The need to improve patient-centered care has led to a growing recognition among clinicians, researchers, and policymakers that clinical trials must include outcomes important to patients, such as quality of life (1). Trials that measure outcomes that are meaningful to patients will ensure that interventions and care address patient needs and priorities. This is critical in the context of CKD, because mortality rates remain high, and the treatment complications and debilitating symptoms can severely impair the patients’ physical and psychosocial wellbeing (2). Fatigue is a frequent, debilitating, and widely experienced symptom of dialysis that profoundly affects patients’ quality of life, but it is rarely measured in clinical trials (2). Of the 1730 hemodialysis trials registered on ClinicalTrials.gov, only 18 (1%) trials included fatigue as an outcome. Similarly, in trials of immunosuppressive agents in kidney transplantation, only 2% included a measure relevant to the assessment of quality of life (3). This mismatch warrants a change in focus of trials—to evaluate outcomes that measure how well patients can live their lives. The increased use of patient-reported outcome measures (PROMs) in trials can achieve this by providing measures of the effect of treatment on quality of life.

The US Food and Drug Administration (FDA) defines a patient-reported outcome as a report that comes from the patient about the status of their health condition (e.g., nausea or pain) without amendment or interpretation of the patient’s response by a clinician or anyone else (4). A PROM provides a quantitative score that reflects how the patient feels or functions in daily life (5). However, selecting a robust and validated PROM from the plethora of available measures is challenging, because the rigor of validation varies, and not all PROMs are feasible to use routinely in clinical research. The specific type of PROM chosen depends on the research question, outcome of interest (global quality of life or specific symptoms, such as fatigue), clinical validation data, and feasibility (e.g., availability of resources). This article will provide a practical outline of the considerations in selecting PROMs and discuss the challenges of using PROMs in clinical trials to assess patient outcomes to provide researchers with a framework for selecting PROMs.

Overview of PROMs

PROMs can be broadly classified into three categories: generic health status, preference based (utility), and symptom- or condition-specific measures. Generic health status measures assess a combination of impairment, disability, and quality of life (6). These are broad in content and relevant to a wide range of patients and conditions. Examples of generic health measures include the 36-Item Short Form Health Survey (SF-36) and the Sickness Impact Profile. Preference-based (utility) measures are also relevant to broad patient populations but allow for a utility value, with a range from zero (dead) to one (full health) to be assigned to the health state described by the patient. The European Quality of Life—Five Dimensions and Health Utilities Index are examples of preference measures. Utility values express quality of life of different health states on a common scale, allowing for comparisons to be made across different populations (e.g., CKD and cancer). In economic evaluations, they can be used to calculate quality-adjusted life-years that provide information about the cost-effectiveness of the intervention. Symptom- or condition-specific measures assess impairment due to specific symptoms experienced by the patient (6). They are used in trials where there is a target symptom or condition, because the broader content of general health questionnaires may not be sensitive enough to capture the unique experiences of patients. The Kidney Disease Quality of Life is an example of a condition-specific measure, which has the same content as the SF-36 but also, includes items that are specific to kidney disease. An example of a symptom-specific measure is the Fatigue Severity Scale.

Selecting PROMs for Clinical Trials in Nephrology

The past three decades have seen a growing number of PROMs. The Patient-Reported Outcome and Quality of Life Instruments Database includes approximately 500 PROMs and provides access to most measures and manuals (7). Another relevant resource is the Patient-Reported Outcomes Measurement Information System (PROMIS), an initiative that develops and validates a set of person-centered measures to evaluate and monitor physical, mental, and social health across various chronic conditions and health statuses. To date, PROMIS has measures for >19 health domains, including fatigue (physical), depressive symptoms (mental), and peer relationships (social) (8).

Several frameworks have been developed to guide the selection of PROMs for clinical research. These...
include guidelines from the FDA for selecting PROMs (of note, the use of validated PROMs in clinical trials is now required by the FDA for the approval of new therapies), the Consensus-Based Standards for the Health Measurement Instruments (COSMIN) checklist for assessing the methodologic quality of studies that examine psychometric properties of measures (9), and the Consensus-Based Standards for the Health Measurement Instruments Outcome Measures in Effectiveness Trials (COMET) guidelines for trialists considering PROMs to include as part of a core outcome set (10). The Consolidated Standards for Reporting Trials Statement now includes an extension for the reporting of patient-reported outcomes, which includes relevant items that need to be addressed when selecting and reporting PROMs for trials (11). We have integrated the aforementioned frameworks and guidance in the following section on principles for selecting PROMs for use in clinical trials (Figure 1).

**Rationale for the Trial**

The selection of a PROM must be considered early in the design of a trial. Initially, the rationale of a trial is informed by the known prevalence and severity of the outcome of interest. Patient input is also needed to ensure that the PROM chosen captures the relevant experiences and assesses aspects of the outcome that are considered important by patients themselves. Patient-reported outcomes may also be identified in core outcome sets—defined as an agreed minimum set of outcomes that should be reported in all trials, because they are explicitly identified to be critically important by all stakeholders (i.e., patients, caregivers, and health professionals). The global Standardized Outcomes in Nephrology initiative was recently formed to establish consensus-based core outcome sets for all stages of CKD, with an initial focus on hemodialysis (Standardized Outcomes in Nephrology Hemodialysis [SONG-HD]). Of note, core outcomes do not have to be included as primary outcomes for a trial. In SONG-HD, fatigue was identified as a core patient-reported outcome to be included in all trials in hemodialysis.

**Research Objectives**

Choosing PROMs must be on the basis of an *a priori* hypothesis of how the intervention may affect patients, which would usually be informed by the evidence from previous trials or observational or qualitative studies (11). For example, in the Frequent Hemodialysis Network trials, health-related quality of life was conceptualized as a multidimensional domain through a review of evidence from previous trials (12). To assess this domain, a comprehensive set of measures was used that comprised a general health status questionnaire (SF-36) and symptom-specific measures, such as the Beck Depression Inventory and the Sleep Problems Index, for a more targeted assessment of the various dimensions of health-related quality of life.

**Evidence of Validity and Reliability**

For a trial, a PROM should have sufficient evidence of psychometric robustness for use in the population of interest. There are several measurement properties that trialists can consider when selecting a PROM (13,14), which we have defined and outlined in the following.

**Reliability**

Reliability is the reproducibility and internal consistency of a measuring instrument (the extent to which the measure is free from random error).

**Internal Consistency.** Internal consistency is the degree of correlation between different items in the measure. There are a number of ways to measure internal consistency of a measure. Cronbach $\alpha$ is commonly used to measure this property, and it is suggested that this coefficient should be between 0.70 and 0.90. Internal consistency can also be assessed by examining the correlation of each item to the measure as a whole by omitting that item. It is recommended that this correlation does not exceed 0.20.

**Reproducibility (Test-Retest Reliability).** Reproducibility (test-retest reliability) is the ability of a measure to yield the same results when repeated. Reproducibility is assessed by examining the degree of agreement between scores on the measure at first assessment and when reassessed. Although some set much higher requirements, commonly accepted minimal standards for this reliability coefficient are 0.70 for group data.

**Validity**

Validity is the degree to which the measure assesses what it claims to assess.

**Criterion Validity.** Criterion validity is the extent to which a measure is related to a gold standard. For PROMs, a perfect gold standard measure against which a brand-new measure can be compared is difficult to find or may not exist. Therefore, it is most often used when a new, shorter version of a measure is being compared with its longer version. Correlation should be high (0.8) in this case.

**Content (and Face) Validity.** Content (and face) validity is the extent to which a measure covers all important dimensions of the health condition to be assessed, and face validity is concerned with whether a measure appears to assess what it intends to assess. Qualitative research to elicit the perspectives and priorities of relevant stakeholders, particularly patients, is typically conducted to establish content validity.

**Construct Validity.** Construct validity is the degree to which a measure assesses the intended outcome (patient-reported validity)

---

**Figure 1.** Outline of the considerations for selecting patient-reported outcome measures for clinical trials.
outcome of interest; e.g., fatigue), which is usually not directly observable. The items on a measure produce a score that theoretically represent this outcome. Construct validity is assessed by examining correlations between these scores and a set of other variables expected to be related (e.g., clinical outcomes). There is no agreed threshold for this correlation, but an appropriate range is 0.40–0.80.

**Convergent and Discriminant Validity.** Both properties are considered subtypes of construct validity and required together as evidence for good construct validity. Convergent validity is the degree to which two measures of an outcome that should theoretically be related are actually observed to be related. Discriminant validity assesses whether measures that are theoretically unrelated are observed to be dissimilar.

**Responsiveness.** Responsiveness is the ability of the measure to detect changes in the outcome over time. Various approaches have been used to assess responsiveness, and the most frequently used statistical indices include responsive ratio (generally accepted ratio of 0.5), effect size (0.8 or greater is considered as large relative size of change), and standardized response mean (similar to effect size but uses SD of change scores as the denominator instead of baseline scores).

**Feasibility and Challenges.** Selecting and implementing PROMs in trials present some challenges related to the response rates, CKD population, assessing symptoms, and limitations of the measures (2). Ensuring high response rates for PROMs is important, because imputation methods for missing data are less robust than would be ideal. Response rates can be influenced by various factors, including feasibility aspects outlined in the COSMIN COMET framework (10): patient’s comprehensibility, acceptability to patients, length of measurement, ease of administration, interpretability of scores, and completion time. Patients with CKD frequently experience an array of debilitating symptoms, including fatigue or pain, increased cognitive impairment, and frailty, and often contend with comorbid conditions (e.g., diabetic retinopathy) that can prevent or diminish their ability to complete PROMs (2), and thus, they may require additional support from clinical or research staff to complete the measure. For patients who are unable to complete PROMs due to specific impairments, proxy-reported outcome measures may be considered, whereby clinicians or caregivers respond on the patient’s behalf (15).

**Conclusion.** The use of appropriate and psychometrically robust PROMs in nephrology trials can provide highly relevant evidence on quality of life and wellbeing of patients to inform shared decision making about treatment in CKD. Despite this, PROMs remain infrequently used in trials in nephrology, and this may be, to some extent, attributed to the challenges and uncertainties in selecting PROMs. Selecting PROMs requires justification on the basis of an informed hypothesis related to the patient population and intervention, explicit identification of its importance to patients, consideration of the relevant measurement properties, and an understanding of the range of potential barriers and feasibility aspects to implementing PROMs. Efforts to improve the uptake of well validated and appropriate PROMs can ensure an accurate and consistent assessment for improving patient-centered trial-based evidence to assess outcomes for patients with CKD.

**Acknowledgments.** The content of this article does not reflect the views or opinions of The American Society of Nephrology (ASN) or the Clinical Journal of the American Society of Nephrology (CJASN). Responsibility for the information and views expressed therein lies entirely with the author(s).

**Disclosures.** A.J. is supported by a National Health and Medical Research Council (NHMRC) program grant BEAT-CKD (APP ID 1092579). A.T. is supported by a NHMRC fellowship (APPID 1108676). The funding organization had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; and preparation, review or approval of the manuscript.

**References.**


Published online ahead of print. Publication date available at www.cjasn.org.