“It’s Valid and Reliable” Is Not Enough: Critical Appraisal of Reporting of Measures in Trials Evaluating Patient Decision Aids

Karen R. Sepucha, PhD, Daniel D. Matlock, MD, MPH, Celia E. Wills, PhD, RN, Mary Ropka, PhD, RN, Natalie Joseph-Williams, GDipPsych, Dawn Stacey, RN, PhD, ChirkJenn Ng, MBBS, PhD, Carrie Levin, PhD, Joanne Lally, Cornelia M. Borkhoff, Richard Thomson, MD

Background. This review systematically appraises the quality of reporting of measures used in trials to evaluate the effectiveness of patient decision aids (PtDAs) and presents recommendations for minimum reporting standards. Methods. We reviewed measures of decision quality and decision process in 86 randomized controlled trials (RCTs) from the 2011 Cochrane Collaboration systematic review of PtDAs. Data on development of the measures, reliability, validity, responsiveness, precision, interpretability, feasibility, and acceptability were independently abstracted by 2 reviewers. Results. Information from 178 instances of use of measures was abstracted. Very few studies reported data on the performance of measures, with reliability (21%) and validity (16%) being the most common. Studies using new measures were less likely to include information about their psychometric performance. The review was limited to reporting of measures in studies included in the Cochrane review and did not consult prior publications. Conclusions. Very little is reported about the development or performance of measures used to evaluate the effectiveness of PtDAs in published trials. Minimum reporting standards are proposed to enable authors to prepare study reports, editors and reviewers to evaluate submitted papers, and readers to appraise published studies. Key words: psychometrics; decision making; decision aids. (Med Decis Making 2014;34:560–566)

The Institute of Medicine (IOM) has identified patient-centered care as central to health care quality, and shared decision making is often discussed as a means to promote more patient-centered care.1,2 Patient decision aids (PtDAs) are evidence-based interventions designed to support shared decision making by preparing patients to participate in treatment decisions. Evaluation of patient-centered interventions, such as PtDAs, requires patient-reported measures, and it is important that these measures have demonstrated strong psychometric performance.

Useful measures should address appropriate or “meaningful” processes and outcomes that are essential to high-quality decision making.3 The International Patient Decision Aids Standards (IPDAS) has recommended 2 core domains for evaluating the effectiveness of PtDAs: decision quality and decision process.4 Decision quality includes subdomains of decision-specific knowledge, realistic expectations, and value concordance (or extent to which treatments match patients’ goals). Decision process measures include subdomains of recognition of a decision, feeling informed about options and outcomes, feeling clear about what matters most, discussing the goals of treatment with providers, and being involved in decision making. The Cochrane Collaboration systematic review has found that PtDAs increase decision quality and improve decision process.5

Previous reviews have evaluated the quality of selected measures used to assess the effectiveness of PtDAs and shared decision making. Kryworuchko and others6 appraised the quality of 8 primary outcome measures used in randomized controlled trials (RCTs) evaluating PtDAs and found that only 2 had evidence of strong psychometric performance. Recently, Scholl and others7 appraised the quality of 19 different measures of shared decision making.
and also found limited reporting on the performance of measures. Without clear data on the performance of measurement instruments, it is difficult for investigators to decide on which measures to use and it is difficult to interpret the results of the studies. Further, the lack of information about performance of measures hampers the ability to generate consensus on a core set of shared measures to facilitate knowledge synthesis and theory building in the field.

This paper extends previous work by conducting a comprehensive review of the measures used to evaluate decision quality or decision process in the RCTs in the 2011 Cochrane review of PtDAs, focusing on the quality of reporting of their development and performance, in order to propose standards to enable authors to prepare their study reports, editors and reviewers to evaluate submitted papers, and readers to appraise published studies.

METHODS

Two reviewers independently reviewed the full-text manuscripts of the 86 RCTs included in the 2011 systematic review of PtDAs\(^5\) and abstracted information using standard forms. The reviewers abstracted information on each time a measure was reported that compared 1 or more aspects of decision quality or decision process for the intervention and control groups. We collected information on study context, description of the measures and their administration, development process, and psychometric performance. Table 1 includes some of the abstracted data fields on the performance of the measures. The online appendix includes the full set of fields. The data abstracted from the studies included in the systematic review are available from the corresponding author by request.

Definitions and examples of items were established prior to data abstraction, and regular discussion among coauthors ensured consistency. A measure was considered new if there was no cited prior publication and/or it was not a known, named scale. Articles that cited a reference with respect to any of these issues—for example, “The DCS has been shown to be valid and reliable (O’Connor, 1998)” —were given credit for reporting those elements. However, we did not consult cited sources to confirm that information or to obtain additional unreported information. Frequent calls with the entire coding group were held throughout the data abstraction process. Discrepancies between reviewers were initially discussed by the 2 reviewers, and the majority were resolved after consulting the full text. Common and unresolved discrepancies were brought for discussion by the entire group, with the lead authors (K.S. and R.T.) adjudicating to ensure consistency across studies and resolve any remaining disagreements. For example, we clarified that the reporting of response rates for the overall study did not provide evidence of acceptability of an individual measure used in the study.

Analysis

We classified the measures and assessed the presence of reporting for key elements of scale development and psychometric performance. We examined reporting for measures of knowledge, values-choice concordance, and decision process. We did not separate out specific elements of decision process (e.g., feel informed), as most measures included multiple elements and did not report separately. We hypothesized that new measures would have more substantial reporting of psychometric performance than previously validated measures; hence, we compared reporting of new to previously published measures. Given that the Decisional Conflict Scale (DCS) was the most extensively used measure, we analyze and
Of the 86 trials in the 2011 Cochrane review, 76 of 86 (88%) measured at least 1 aspect of decision quality or decision process. Most of the remaining 10 studies (7/10, 70%) evaluated the impact of decision aid on choices or uptake of tests or treatments.

Across all studies, we abstracted 178 instances where a measure of 1 or more aspects of decision quality and/or decision-making process was reported. Of the 178, 73 (41%) were related to knowledge and/or realistic expectations, 13 (7%) covered value-choice concordance, and 92 (52%) covered 1 or more aspects of decision process. The following results summarize the reporting on the development process for the knowledge, value-choice concordance, and decision process measures. Table 2 presents details about the frequency and type of information reported about the 178 abstracted measures, comparing new and established measures.

Sixty studies included 73 instances of use of measures of knowledge and/or realistic expectations. Only 2 of the knowledge measures were named, so it was difficult to ascertain whether the same measures were used across studies. About half (41/73, 56%) appeared to be new measures used for the first time in the study. Thirty-five knowledge measure descriptions (48%) either included information on item generation (13/73) or referenced a prior publication for item development (3/13).

Sixty-one studies included 92 instances of measures of decision process. Most (78/92; 85%) were
named, established measures, while 14/92 (15%) were developed specifically for that study. About half (47/92, 51%) were some version of the DCS. None included data on item generation in the paper itself, although about half (50/92, 54%) referenced a prior publication. Not all studies cited prior work when using a named scale; for example, 15 studies using the DCS did not include a citation for it.

DISCUSSION

Decision quality and decision process have been identified as core measures to evaluate the effectiveness of PtDAs.\(^4\) This review of the 86 RCTs of PtDAs found that reporting on the development and performance of measures was extremely limited. About one-third of the studies used new instruments, and, contrary to our expectation, the amount reported on new measures was actually less than that reported on existing measures.

For the measures included in the review, reliability and validity were most commonly reported aspects of performance, but only for 21% and 16% of instances, respectively. In several instances, this was limited to statements such as “the scale is valid and reliable” with a reference to a prior citation. Other important features such as pilot testing, responsiveness, precision, and acceptability were reported in less than 10% of instances, and virtually all of those reports were from the DCS. A particular gap was found with cognitive interviews and feasibility, which none of the studies mentioned.

Studies should include relevant details on the psychometric properties of measures used, so that readers can appropriately interpret results and conclusions. Psychometric performance can vary across settings, samples, and measurement contexts.\(^8\)–\(^10\) A common misperception is that if one simply picks a validated and reliable survey instrument, then there is no more work to be done. In reality, validity and reliability are not properties of a survey instrument; rather, the data and the interpretation of the data (which includes understanding the administration, setting, sample, and analysis procedures) support construct validity.\(^10\) Thus, relevant information on psychometric performance ideally needs to be reported for each study and each use of an instrument or measure.

To support researchers in reporting the development and performance of measures, we outline a set of reporting standards for both new and existing measures used to evaluate PtDAs in Table 3. The recommended standards are applicable for any PtDA evaluation measure. The first 5 items are basic information required for a reader to understand what data were collected and how they were analyzed. For decision aid studies, it is important to understand the context and timing of the assessment (such as how long after the decision aid the measure was administered, whether a consultation with a provider occurred, whether treatment was completed) and the mode of administration. Even with respect to these basic issues, information was often lacking in the reviewed studies.

Items 6–8 in Table 3 focus on the performance of the measure, and for each of these items there is

---

Table 2 Reporting on Performance of New and Established Measures of Decision Quality and Decision Process in Studies of PtDAs

<table>
<thead>
<tr>
<th>Item generation(^b)</th>
<th>New Measures (n = 61)</th>
<th>DCS(^a) (n = 47)</th>
<th>Non-DCS (n = 70)</th>
<th>All Measures (N = 178)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20%</td>
<td>68%</td>
<td>66%</td>
<td>51%</td>
</tr>
<tr>
<td>Cognitive testing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pilot studies(^b)</td>
<td>11%</td>
<td>0</td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td>Reliability</td>
<td>15%</td>
<td>30%</td>
<td>21%</td>
<td>21%</td>
</tr>
<tr>
<td>Validity</td>
<td>15%</td>
<td>17%</td>
<td>17%</td>
<td>16%</td>
</tr>
<tr>
<td>Responsiveness(^c)</td>
<td>0%</td>
<td>19%</td>
<td>0%</td>
<td>5%</td>
</tr>
<tr>
<td>Accuracy/precision</td>
<td>3%</td>
<td>15%</td>
<td>0%</td>
<td>5%</td>
</tr>
<tr>
<td>Interpretability(^c)</td>
<td>0%</td>
<td>13%</td>
<td>1%</td>
<td>4%</td>
</tr>
<tr>
<td>Acceptability</td>
<td>2%</td>
<td>4%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td>Feasibility of administration</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: DCS = Decisional Conflict Scale; PtDAs = patient decision aids.  
\(a\). Includes all versions of the DCS as well as studies that only included some of the DCS subscales.  
\(b\). \(P < 0.01\) comparing frequency of reporting on new and prior published measures.  
\(c\). \(P < 0.05\) comparing frequency of reporting on new and prior published measures.
considerable discretion as to the amount of detail to include. A full assessment of these items could provide enough material for an entire manuscript, or even a book, in the case of well-tested measures. The next 3 paragraphs provide some guidance on what would satisfy a minimum reporting requirement and then what would be considered a strong or enhanced reporting for these 3 items.

Some information on current experience with a measure should be included. All studies using new or previously published instruments will have data to report on some psychometric assessments, such as internal consistency reliability, acceptability, and even validity. Not all of these elements are required for each measure, and the type of information most relevant will vary depending on the measure. For studies reporting knowledge using a new measure, it is more important to provide some evidence that the measure demonstrates discriminant validity than internal consistency. The example in Table 3 taken from Barry and others\textsuperscript{11} illustrates how they demonstrated discriminant validity by comparing responses of experts and patients and found significant differences in knowledge scores. In contrast, studies using the total score from the DCS should report on the internal consistency in

\begin{table}[h]
\centering
\small
\begin{tabular}{|l|l|}
\hline
Type of Measure & \\
\hline
\textit{Previously published measures} & \\
1) Name and version of instrument & Extract from Vodermaier and others\textsuperscript{12} \textit{Primary outcome variable} \textit{Decisional conflict. The Decisional Conflict Scale (DCS; O’Connor, 1995) measures patients’ uncertainty about which treatment to choose, factors contributing to uncertainty (believing to be uninformed, unclear values, and unsupported in decision making), and perceived effectiveness of decision making. Questions have to be answered on a 5-point Likert scale [from strongly agree to strongly disagree]. Higher scores on the scale or subscales reflect higher decisional conflict, uncertainty, and a less effective choice. The German version of the scale demonstrated subscale and total score internal consistencies in the present sample between 0.73 and 0.94. The scale discriminates between patients who make and those who delay decisions (O’Connor, 1995; Bunn and O’Connor, 1996) and is sensitive to change (O’Connor et al., 1998).} \\
2) Domains measured & \\
3) Mode and timing of administration & \\
4) Number of items and response format & \\
5) How the measure is scored (including missing items if applicable) & \\
6) Experience with the measure in the current study (e.g., reliability, validity, acceptability) & \\
7) Type of psychometric performance previously published & \\
8) Adequacy and relevance of prior performance to current samples and settings & \\
\hline
\textit{New measures} & \\
Newly developed measures should describe items 1–6 above and should include the following: & \\
1) Measure development with a focus on content validity and understandability & Extracts from Barry and others\textsuperscript{11} \textit{Validation of new outcome measures} \textit{Cronbach’s alpha statistic for the items testing BPH knowledge was 0.68. The criterion validity of this test was assessed by comparing scores for a convenience sample of 12 urologic nurses with the scores of the 167 BPH patients enrolled in the baseline period. The nurses had a mean score of 14.8 [out of 20], compared to 5.6 for the patients (p < 0.001). Nurses answered an average of 85% of the questions correctly, compared to 48% for the patients (p < 0.001). Furthermore, a modest correlation between these patients’ knowledge scores and their educational levels was seen, r = 0.23 (p < 0.001).} \\
2) If permitted by publisher, include the entire measure in the paper or as an appendix; otherwise, state how the measure can be accessed & \\
\hline
\end{tabular}
\caption{Proposed Minimum Reporting Standards for Measures Used to Evaluate the Effectiveness of Patient Decision Aids (PtDAs)}
\end{table}

BPH = benign prostatic hyperplasia.
that sample and may refer to prior citations for evidence of validity, as illustrated in Table 3 from Vodermaier and others. ¹²

Often, prior publications can be cited to provide evidence of psychometric properties. When including a citation, at a minimum one should provide specific acknowledgement of the kind of information that is included in the citation (e.g., development process, reliability, acceptability). Even better would be to provide some details on the strength of performance of the measure, such as including an internal consistency reliability coefficient to describe reliability, naming the other measures used to establish divergent and/or convergent validity, and presenting the magnitude of the associations. Several textbooks have detailed information on how to assess the adequacy of the psychometric evidence; one that we recommend is by Waltz and others.¹³ Finally, an overall assessment should occur, perhaps in the discussion, that critically reflects on the performance of the measure in the current study and how aspects of the sample, administration, or scoring extend the known properties of the measure.

In some situations, a measure may not exist that is appropriate for a given study. For decision aid studies, this situation often arises with knowledge measures and measures of patients’ goals and preferences, which are specific to the decision. Researchers should be aware of several resources for potential measures, such as the National Cancer Institute’s GEM (www.gem-beta.org), the Ottawa Health Research Institute’s evaluation section (http://decisionaid.ohri.ca/eval.html), and the National Institute of Health’s PROMIS (www.nihpromis.org). Sometimes it may be preferable or necessary to develop a new measure as opposed to adapting an existing one. Often, these measures are designed and used in one study, and the authors may not have any plans to use the measure again or to make it widely available. The limited use, however, does not mean that it is not important to develop a strong measure, particularly if it is a main outcome measure for a study of an intervention. As shown in Table 3, reporting on newly developed measures requires the basic information (Items 1–5) as well as details on the development. In particular, it is necessary at a minimum to include some information on content validity, or how the items were generated (e.g., Was the approach driven by theory? How much input did patients have? Was there any empirical testing?), and on how the developer ensured understandability (e.g., focus group testing, cognitive interviewing, pilot testing). Additional details about the development process are desirable but often are not feasible to include in the manuscript itself. A link to an online supplement or reference to another publication where readers may obtain additional details is helpful. Finally, for new measures, it is important to provide the items or access to the items in the manuscript, in an online appendix, or on a Web site.

Other organizations have promoted increased attention to measurement properties and have made recommendations for patient-reported outcome measures in treatment trials.¹⁴ The minimum standards proposed here are aligned with broader guidelines and adapted for measures of decision aids. The purpose of this effort is to support researchers in selecting measures and enabling more complete reporting within the scientific literature. However, tensions exist in both the reporting and the development of measures, which must be acknowledged. With regard to reporting, there is a practical tension around how much can reasonably be reported within a manuscript. Given the word limits imposed by many journals, it may be necessary to limit the amount of detail reported on psychometrics in the manuscript. However, many journals will allow authors to include online supplements that may alleviate the restrictions on word limits and make this information available to readers, reviewers, and editors. Finally, measure developers can post user guides on their Web sites or on a centralized Web site such as the National Cancer Institute’s Grid Enabled Measures project (see, e.g., www.gem-beta.org, http://decisionaid.ohri.ca/eval.html, or http://www.massgeneral.org/decisionsciences/).

A second tension exists with regard to balancing the requirements for strong psychometrics with the need to develop new measures. If every measure has to be extensively examined prior to its use in a study, this could have the unintended consequence of inhibiting innovation and/or slowing the pace of research. Researchers may avoid creating new measures, even in situations where no measure exists. Overreliance on existing measures may prevent testing of new theories or hypotheses. The standards outlined in Table 3 can be achieved with a reasonable amount of effort, particularly if researchers are thinking about evaluation early in the process. For instance, it would not require much additional effort to develop and test decision-specific measures of knowledge or goals in conjunction with the development of a decision aid.

There is a benefit to having a core set of measures that are used consistently across studies. If every investigator creates his or her own measures, then
cross-study comparisons become impossible and the field as a whole will have a more difficult time moving forward. To help bring standardization to the field, a new resource is available for formal and informal reporting on measure constructs and instruments through the National Cancer Institute’s Grid Enabled Measures shared decision-making database. Ideally, this effort will support researchers in selecting strong measures and enable more complete reporting.\textsuperscript{15}

Our study has several limitations. First, we only included RCTs from the Cochrane review of PtDAs. Second, we did not review the cited sources for previously published measures, as we were focused on the quality of reporting, not the quality of the measures themselves. As a result, these findings should not be interpreted to mean that the measures used were necessarily poor; rather, the reporting of measures was inadequate.

While not every psychometric property should be reported for every measure in every publication, current practice is clearly inadequate. Improving the process and quality of decision making for patients is of paramount importance; accurate measurement and reporting of the performance of the measures used are essential to moving the field forward.

REFERENCES