VALUE IN HEALTH 17 (2014) 501–516

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

ISPOR TASK FORCE REPORT

PRO Data Collection in Clinical Trials Using Mixed Modes: Report of the ISPOR PRO Mixed Modes Good Research Practices Task Force

Sonya Eremenco, MA[1,], Stephen Joel Coons, PhD[2], Jean Paty, PhD[3], Karin Coyne, PhD[1], Antonia V. Bennett, PhD[4], Damian McEntegart, BSc[5], on behalf of the ISPOR PRO Mixed Modes Task Force*

[1]*Outcomes Research, Evidera, Inc., Bethesda, MD, USA;* [2]*Patient-Reported Outcome Consortium, Critical Path Institute, Tucson, AZ, USA;* [3]*Endpoint Strategy, Quintiles, Hawthorne, NY, USA;* [4]*Department of Health Policy and Management, University of North Carolina, Chapel Hill, NC, USA;* [5]*Consultant, Nottingham, UK*

ABSTRACT

The objective of this report was to address the use and mixing of data collection modes within and between trials in which patient-reported outcome (PRO) end points are intended to be used to support medical product labeling. The report first addresses the factors that should be considered when selecting a mode or modes of PRO data collection in a clinical trial, which is often when mixing is first considered. Next, a summary of how to "faithfully" migrate instruments is presented followed by a section on qualitative and quantitative study designs used to evaluate measurement equivalence of the new and original modes of data collection. Finally, the report discusses a number of issues that must be taken into account when mixing modes is deemed necessary or unavoidable within or between trials, including considerations of the risk of mixing at different levels within a clinical trial program and mixing between different types of platforms. In the absence of documented evidence of measurement equivalence, it is strongly recommended that a quantitative equivalence study be conducted before mixing modes in a trial to ensure that sufficient equivalence can be demonstrated to have confidence in pooling PRO data collected by the different modes. However, we also strongly discourage the mixing of paper and electronic field-based instruments and suggest that *mixing of electronic modes be considered for clinical trials and only after equivalence has been established.* If proceeding with mixing modes, it is important to implement data collection carefully in the trial itself in a planned manner at the country level or higher and minimize ad hoc mixing by sites or individual subjects. Finally, when mixing occurs, it must be addressed in the statistical analysis plan for the trial and the ability to pool the data must be evaluated to then evaluate treatment effects with mixed modes data. A successful mixed modes trial requires a "faithful migration," measurement equivalence established between modes, and carefully planned implementation to minimize the risk of increased measurement error impacting the power of the trial to detect a treatment effect.

Keywords: electronic PRO, ePRO, equivalence, mixed modes.

Introduction

The incorporation of the patient perspective in the evaluation of medical products (i.e., drugs, biologicals, and devices) is increasingly important and considered essential in many cases. Medical products aimed at relieving patients' symptoms and/or improving levels of self-reported functioning will require measures of patient-reported outcomes (PROs) as end points in clinical trials. A PRO instrument systematically collects treatment benefit data directly from patients, without interpretation by clinicians or others [1]. As stated in the 2009 US Food and Drug Administration (FDA) Guidance for Industry titled Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims ("PRO Guidance"), "Use of a PRO instrument is advised when measuring a concept best known by the patient or best measured from the patient perspective" [1].

There is no doubt that the release of the FDA's PRO Guidance has focused increased attention on the development and use of scientifically sound measurement of PRO end points in clinical trials. In addition, clinician-reported outcome (ClinRO) measures, observer-reported outcome (ObsRO) measures, and performance outcome (PerfO) measures are receiving increasing attention [2,3]. ClinRO measures are completed by clinicians and are often based on clinical interviews (e.g., Hamilton Depression Rating Scale in depression trials). ObsRO measures are completed by nonclinical informants (e.g., spouse, caregiver, parent, or teacher) and report on observable disease- and/or treatment-related concepts (e.g., activities of daily living inventory completed by caregivers

* *Address correspondence to:* Sonya Eremenco, MA, Evidera, Inc., 7101 Wisconsin Avenue, Suite 600, Bethesda, MD 20814.
E-mail: sonya.eremenco@evidera.com.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2014.06.005

Background to the Task Force

In February 2011, the ISPOR Health Science Policy Council recommended to the ISPOR Board of Directors that an ISPOR Good Research Practices Patient-Reported Outcomes (PRO) Task Force be established to focus on mixed modes of PRO data collection and to provide good research practice recommendations for the analysis of mixed modality data. These recommendations were intended to address the use of data from multiple data collection modes raised in the 2009 US Food and Drug Administration (FDA) Guidance for Industry titled Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [1], which the prior ePRO Task Force [29] did not address. The Board of Directors approved this PRO Task Force in March 2011.

Members and primary reviewers were selected to represent a diverse range of perspectives, including government (FDA), academia, research organizations, and the pharmaceutical industry. The task force leadership group was comprised of experts in PRO assessment, clinical trial data collection, regulatory affairs as well as design and development of ePRO technology. In addition, the task force had international representation with members from Canada, Singapore, and the United Kingdom in addition to the United States.

The Task Force met approximately every five weeks by teleconference to develop an outline and discuss issues to be included in the report. In addition, task force members met in person at ISPOR International Meetings and European Congresses. All task force members, as well as primary reviewers, reviewed many drafts of the report and provided frequent feedback in both oral and written comments.

Preliminary findings and recommendations were presented four times in forum and workshop presentations at the ISPOR Annual European Congresses in 2011 and 2012 as well as the ISPOR Annual International Meetings in May 2012 and 2013. Comments received during these presentations were addressed in subsequent drafts of the report. In addition, the draft task force report was sent out to the nearly 500-person ISPOR PRO Review Group twice.

All comments were considered, and most were substantive and constructive. The comments were discussed by the task force in a series of teleconferences and addressed as appropriate in revised drafts of the report. All written comments are published at the ISPOR website on the task force's webpage: http://www.ispor.org/TaskForces/PRO-mixed-modes.asp. The task force report and webpage may also be accessed from the ISPOR homepage (www.ispor.org) via the purple Research Tools menu, ISPOR Good Practices for Outcomes Research, heading: Patient Reported Outcomes & Clinician Reported Outcomes heading and link: http://www.ispor.org/workpaper/practices_index.asp. A list of leadership group members is also available via the task force's webpage.

Once consensus was reached by all task force members, the final report was submitted to *Value in Health* in June 2014.

in Alzheimer's disease trials). PerfO measures are assessments based on a task performed by a patient according to instructions administered by a health care professional, and rely on the cooperation, ability, and motivation of the subject.

Although the final PRO Guidance addressed only PRO instruments when it was released in 2009, the FDA held a public workshop in October 2011 in which it discussed the need for the same level of evidence (i.e., well-defined and reliable measurement) for all clinical outcome assessment tools (i.e., PRO, ClinRO, ObsRO, and PerfO measures) intended to support medical product labeling claims [2,4]. Similarly, many of the recommendations in this task force report apply to ClinRO and ObsRO measures as well, but the focus is on PRO measures.

In addition to the FDA's increased focus on well-defined and reliable assessment of clinical trial end points, one of the most important developments in the field of PRO measurement has been the emergence of technologies that enable the collection of data electronically. Advantages of using electronic data collection include less subject burden, avoidance of secondary data entry errors, easier implementation of skip patterns, date and time stamping, reminders/alerts, edit checks, and more accurate and complete data [5–14]. With the increasing availability of multiple modes of PRO data collection, including paper and various electronic formats, the opportunity exists to mix these modes within and across clinical trials in a medical product development program. (Before proceeding further, a clarification regarding terminology is in order. It should be noted that the term *mode of data collection* as used in this report differs from the FDA's terminology. The PRO Guidance makes a distinction between PRO instrument administration modes and data collection methods. According to the PRO Guidance, administration *mode* refers to self- versus interviewer-administered PRO measurement, whereas data collection *method* refers to the tool used for capturing the data such as paper-based questionnaires, Web-based data entry, interactive voice response (IVR) system, or any other ePRO devices [1]. [*Note*. An interviewer-administered PRO measure is not a ClinRO measure because the patient's responses are not interpreted, but simply recorded, by the interviewer.] We find that the distinction made by the FDA is potentially misleading because the term *mixed methods* in the larger PRO measurement field refers to mixing qualitative and quantitative methods in research. This term is *not* associated with multiple methods of data collection. However, the PRO measurement field has a long history of using the term *mixed modes* to refer to both administration and data collection (i.e., ePRO vs. paper). Therefore, to simplify the discussion in this report, we use the term *mode* in the context of both modes of administration per the PRO Guidance and modes of data collection per the PRO measurement field.)

Although mixing of modes within and across clinical trials may meet the needs of global product development programs in which the patient population and access to technology vary considerably within and across regions, such mixing may in fact be an *avoidable source of measurement error*. It is, therefore, the general recommendation of this ISPOR PRO task force report that PRO data collection modes *not be varied within a single clinical trial or between trials that seek to pool or compare the data* without prior evidence of sufficient measurement equivalence between the modes.

This general recommendation is based on the basic research design tenet that anything with the potential to introduce measurement error into a trial should be avoided [15]. Measurement error is, in essence, noise (error variance) that reduces statistical power and attenuates the ability of the trial to detect real change (i.e., treatment effect) in the trial end point. In the context of collecting PRO data—where patients are providing information directly—there are many unavoidable sources of measurement error, including differences introduced by the need to translate and culturally adapt multiple versions of a PRO instrument, specific cultural biases introduced by differing experiences of the medical condition being studied [16], and the variability in subjects' ability to reflect and provide a response.

Potential error variance can also be introduced into the trial design by different data collection modes used within the trial that do not provide comparable data (i.e., the modes lack sufficient measurement equivalence). Because the mode of PRO data collection is a part of the research design, it should be possible (even though challenging at times) to decide on and

deploy a single consistent mode of PRO data collection in the trial. The recommendation of this task force report is to avoid, where possible, *all* potential sources of measurement error, including mixed modes of PRO data collection.

However, mixing of PRO data collection modes within trials does occur and has to be addressed pragmatically. When modes have been directly compared in cross-sectional studies, there is evidence that PRO data collected electronically can be comparable to that obtained by paper-based data collection, particularly with screen-based devices [17]. The literature, however, is not definitive and can be limited by selective reporting; it has been well documented that studies with positive findings are more likely to be published than those with inconclusive or negative results [18,19]. In addition, comparability of data collected on different modes is likely dependent on the specific PRO measure being used; hence, a general assumption of measurement equivalence between or among modes may not always hold. Although some evidence has shown comparability between paper and visual modes [17,20] or between paper, Web-based, and/or IVR modes [21–26], other studies have shown a tendency toward systematically lower scores on electronic versions than on paper versions despite evidence of equivalence between the modes [27,28]. Thus, more evidence is needed to support mixing modes within a trial setting to ensure that it has minimal impact on the results of the trial, particularly when one mode is paper. Nevertheless, this task force does not rule out the possibility that at some point in the future, sufficient evidence will be available to support the assumption of measurement equivalence across modes in most circumstances in which a faithful migration has occurred.

Furthermore, it is clear that this issue of mixing modes was contemplated by FDA in the development of its PRO Guidance. Specifically, the PRO Guidance states that "We intend to review the comparability of data obtained when using multiple data collection methods or administration modes within a single clinical trial to determine whether the treatment effect varies by methods or modes" [1]. The PRO Guidance does not, however, discuss ways for clinical trial designs to ensure the comparability of the data when mixed modes are used. To our knowledge, there are no European Medicines Agency (EMA) documents that speak to this issue. Although this task force report specifically addresses multiple modes of PRO data collection, it is assumed that many of the same issues are involved when multiple modes of ClinRO and ObsRO data collection are considered. Therefore, this report addresses key issues in mixing modes, specifically focusing on how to reduce the impact on measurement error when these situations arise.

The launching point for this task force report is the previous ISPOR electronic PRO (ePRO) task force report by Coons et al. [29], which addressed the evidence needed to support measurement equivalence when migrating from paper to electronic modes of PRO data collection. According to that report,

> measurement equivalence is a function of the comparability of the psychometric properties of the data obtained via the original and adapted administration mode. This comparability is driven by the amount of modification to the content and format of the original paper PRO questionnaire required during the migration process. The magnitude of a particular modification is defined with reference to its potential effect on the content, meaning, or interpretation of the measure's items and/or scales. [29]

Thus, establishing measurement equivalence is essential in demonstrating that the migration from paper to electronic, or for that matter from any data collection mode to another, did *not* affect the instrument's meaning, interpretation, and resulting responses. In the context of the current task force report, we use the term "measurement equivalence" to emphasize the need for the instrument to be *measuring* the same thing regardless of the mode.

Coons et al. [29] did not address the issues to take into account when considering mixing two or more modes of PRO data collection in a single trial or across trials intended to be compared or pooled. This current report builds on the recommendations for changing modes of administration in the original ISPOR ePRO task force report by providing additional recommendations regarding good research practices for migration across modes of data collection and an in-depth exploration of the assessment of measurement equivalence between original and migrated versions of PRO instruments, particularly in the context of mixing data collection modes. In addition, we discuss issues that must be considered to avoid sources of measurement error that materially affect the meaning and interpretation, and consequently the measurement properties, of the instrument being used to assess PRO end points in clinical trials. The report concludes with recommendations for operational and statistical considerations when modes are mixed in a clinical trial setting. The overall objective of the report was to address the use and mixing of data collection modes within and between trials in which the PRO end points are intended to be used to support medical product labeling.

Process for Selecting the Appropriate Mode of Data Collection

The emergence of new technologies allows trial protocols to be written in which data collection schedules and locations can support more timely and convenient assessment of end points. Selecting the appropriate mode of data collection is essential to the success of the trial to ensure that the mode is suited to the trial, population, and PRO measure. The mode selection process may lead to a consideration of mixed modes because of the realization that many modes are available and potentially suitable. See Appendix A in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.06.005 for an overview of common modes of PRO data collection used in clinical trials.

The selection of a PRO-based clinical trial end point measure and the mode of PRO data collection should *not* be an afterthought in a drug development program. All too often, the PRO data collection mode appears to be given insufficient attention, with providers of ePRO technologies and services asked to accomplish the near impossible before the launch of a trial. Hence, as early as possible, a substantial amount of thought and deliberation should be invested in the selection and evaluation of the mode of PRO data collection for a clinical trial. A lead time of 6 months is ideal because ePRO development activities are front-loaded and must occur before the launch of the trial. There are a number of factors that must be considered, including patient population, characteristics of the instrument (e.g., length and format of responses/answers), location of data collection, data collection schedule (which is driven by the type of outcome being assessed), feasibility, and cost.

Patient Population

The primary consideration for the selection of a PRO data collection mode is the patient population that will be asked to provide the self-reported data. The characteristics of clinical trial subjects, particularly sensory and physical abilities, will be important drivers of the choice of modes. It should be noted that this is not a new consideration. Historically, when data collection mode options were limited to paper and pencil, or an interviewer reading to the patient, we had a limited ability to respond to the variability in patients' capabilities. Given the diversity of options now available, we can (and should) be more responsive to the

patient population's needs. For example, subjects who have noncorrectable visual or hearing impairments will require an auditory or visual-based data collection system, respectively. Furthermore, in conditions in which there are decrements in physical function (e.g., joint stiffness and tremors) or patients' physical abilities are compromised, such as rheumatoid arthritis or Parkinson's disease, both the selection and the specifics of the data collection mode will be important. Auditory systems may be good for such patients, or visual systems that have larger font sizes for reading or larger stylus sizes for arthritis sufferers would be helpful.

The Characteristics of the PRO Instrument

The characteristics of the PRO questionnaire can be a critical driver in selecting the mode of data collection. With regard to the length of the PRO instrument, both the number of items and the amount of time necessary to complete the items should be considered. It should be noted that subject burden is an issue to bear in mind regardless of the data collection mode. Raymond [30] makes the distinction between "questionnaires" and "diary-type reports," with the latter comprising fewer concepts with questions that are completed at least daily. Handheld devices have become the mainstay of field-based data collection (eDiaries) in clinical trials, but they are less than optimal for longer PRO instruments. Long and time-consuming questionnaires can be physically and/or cognitively fatiguing and should be avoided regardless of the mode of data collection. Likewise, there may be aspects of the data collection mode to consider that may mitigate or aggravate the fatigue factor.

Depending on the length or complexity of the response options, screen size can be a limitation. If information on a screen is needed to inform or interpret the task or content on a *subsequent* screen, the implementation of the PRO instrument is far from optimal [31]. Hence, the response options should appear on the same screen as the question. Scrolling to access response options should not be required. In the context of IVR systems, memory may be required to enable the subject to select among response options; numerous or lengthy response options complicate task completion. In addition, there are some types of response formats that are not easily operationalized on all data collection platforms. For instance, a traditional visual analog scale (VAS), which is a line with descriptive anchors at each end (e.g., "No pain" to "Pain as bad as it could be") with no intermediate positions along the continuum, does not lend itself to administration on an auditory system (e.g., IVR). Open-ended or free-text responses tend to be more burdensome on tablets and handheld devices because an onscreen keyboard is required for text entry.

Location of Data Collection

In a clinical trial, PRO data may be collected from subjects at the investigative site (e.g., clinic) (hereafter, site-based), in the field (i.e., away from the study site such as subject's home or workplace) (hereafter, field-based), or both. At the study site, the portability of the data collection mode is not as critical; hence, all modes are potentially viable. If the data collection takes place in the field, then subject convenience and portability are important considerations.

Data Collection Schedule

The frequency of protocol-driven data collection points in a clinical trial should also be considered. Technology has enabled greater flexibility and functionality for designing trials with more frequent data capture. Trial protocols may require that PRO data be collected monthly, weekly, daily, or multiple times per day. Within a day, data capture can be scheduled for specific times (e.g., 7 AM and 7 PM) or based on an event (e.g., bowel movement) or symptom (e.g., pain). The choice of mode must consider the multiple places where the subject may be when the data collection is to occur. Hence, for multiple data collection points during the day, the portability of the device is a major consideration assuming that the patient is mobile.

Feasibility of Implementation

Another consideration for mode selection is the infrastructure available in the selected locale of an investigative site or in the trial more broadly. Some regions within a country and some countries more broadly may not be able to support certain technologies. For example, if high-speed Internet access or a cellular phone network is critical for a data collection technology, this will be a criterion for selecting this mode in a trial or at a specific site. Variability in feasibility of implementation across investigative sites within a trial can potentially lead to mixing of modes.

Cost

The reality of conducting clinical trials is that cost will be a factor in determining the mode of data collection. Although this is less than ideal with regard to choosing the optimal mode to collect high-quality data, cost is, nonetheless, a significant driver for selecting a mode. The team making the mode of data collection decision will need to balance the above selection criteria against available funds to make the best decision for a specific circumstance.

It should be noted that if a less expensive option is chosen early in a medical product's development, the team will need to consider using a different mode (i.e., mixing modes) later in the program. *It is likely that choosing one mode of data collection will be more cost-effective than choosing multiple modes in the same trial.*

Furthermore, it cannot be assumed that the paper mode of data collection will always be less expensive than an electronic mode; there are hidden costs with paper, including the time required for secondary data entry into the system and the time spent on "data cleaning" and queries before database lock. Therefore, when cost comparisons are made, the *full* cost of a data collection mode(s) accrued across the lifetime of the study—not just the upfront costs associated with ePRO implementations—should be taken into consideration. Finally, it should be noted that costs associated with electronic modes of data collection change as technology evolves.

After the appropriate mode of data collection has been selected, if the PRO instrument has not previously been implemented in the chosen mode, the next step in the process is to migrate the PRO instrument to the newly selected mode before implementing PRO measurement in the planned clinical trial program. In some instances, multiple modes may be selected for data collection and migration may occur concurrently. *Following migration, it is important to assess whether the new mode has measurement equivalence with the original mode,* that is, *subjects interpret and respond to the instrument the same way regardless of the mode.*

Migration

The discussion of the optimal migration process is relevant to the discussion of potentially using mixed modes because it is necessary to migrate and evaluate measurement equivalence *before* mixing so that the desired modes are available and appropriate for use. Because there is little, if any, literature available on the migration process, these recommended good practices are based on previous successful migrations conducted

by members of the task force that led to demonstrations of measurement equivalence.

The goal of any migration is to have minimal or no impact on the measurement characteristics of the instrument. A "faithful migration" refers to the development of alternative modes of data collection that do not introduce response bias that results from changes in the way the instrument is presented/formatted or how the subject interacts with it. The most common path is migrating from paper to electronic modes, but migration from one electronic mode to another or from electronic to paper also occurs. The primary goal of the migration process is to ensure that subjects interpret and respond to the questions/items on the PRO instrument the same way regardless of the data collection mode. It is possible to evaluate this by conducting cognitive interviews with subjects from the target population and/or assessing response equivalence between modes.

Furthermore, it is possible to achieve measurement equivalence even if the instructions or item presentation may not be the same as the original mode. In fact, there may be an opportunity to present items or instructions within the instrument more clearly in a specific data collection mode. For example, if the instrument has a skip pattern, the electronic version could show only the items that subjects need to complete; that is, if there are any skips or jumps over items based on a previous answer, the subject will never see the item. This can prevent completion of a nonrelevant item and resulting data entry queries due to conflicting responses. Such enhancements may bring clarity to instrument completion, but leaves the instrument "faithful" to its original intent and meaning.

This section on migration issues builds on the recommendations of the previous ISPOR ePRO Task Force report [29] to provide more detailed guidelines for the "faithful migration" process and to discuss the mixing of data collection modes, particularly the special considerations when mixing paper and electronic modes.

Perform a Faithful Migration

Most migrations involve making changes to the PRO instrument that are required because of characteristics of the new mode. A faithful migration is conducted carefully to ensure that only necessary changes to the format and instructions are made—item and response content has *not* changed. A faithful migration of an instrument does *not* need to look exactly like it did originally, but it does need to collect the same data.

The degree of modification is a key consideration in determining the level of evidence needed to evaluate equivalence as presented by Coons et al. [29], and is a direct result of the migration process. For example, migrating from paper to an IVR presentation has been categorized as a moderate modification [29] because the necessary changes are more extensive than with most other data collection modes. Modification includes revisions to the instructions and may include nonsubstantive changes to the wording of questions and responses for effective implementation on the IVR system. These changes, along with the change from visual to auditory cognitive processing, may result in systematic differences in responses between the modes [29]. It should be noted, however, that traditional telephone-based data collection using a live interviewer rather than the recorded scripts of an IVR system would require the same type of changes.

The recommended steps to conduct a "faithful" migration are discussed in detail in Appendix B in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.06.005. Further cognitive interviewing, usability testing, and/or equivalence testing may be required to confirm that the migration has been faithful and the new implementation is capturing the same data as the original. These are addressed in later sections of this report.

How much of the time during the **past 4 weeks** . . .	None of the time	A little of the time	Some of the time	Much of the time	All of the time
1. Did you feel worn out?	0	1	2	3	4

Fig. 1 – Example of item 1 in original paper format.

Mode-Specific Considerations for Migration

In addition to the process outlined in Appendix B, each mode has specific considerations that must be addressed during the migration process. The four most common modes will be addressed here.

Migration to a smartphone/handheld device

The main factor for consideration in migrating to a smartphone/handheld device is the space constraints of the smaller screen. Regardless of how the instrument was originally formatted on paper, the default on a handheld device is one item per screen, with all responses visible on the same screen. In some cases, a long item with long response options will pose a great challenge because it is not possible to fit all the text of the question and responses on the screen at the same time. Some solutions are to display the item on the first screen and responses on the second screen with a reminder of what the question was asking, or to present partial responses on a line with a popup that displays the entire response for clarification.

Another possibility is to allow scrolling on the screen to accommodate longer questions or to view response options that do not fit on the small screen. Scrolling is not ideal because it greatly increases the risk that subjects will not be aware that it is necessary to scroll to view the missing text or response options. It also increases the risk of subjects interpreting the question differently or answering differently.

Figures 1 and 2 illustrate the paper to electronic migration process to show a number of changes that were necessary to accommodate the item in the ePRO version. As mentioned in Appendix B, migration from some paper-based questionnaires will require the rejoining of split item stems. In Figure 1, the paper version presents the item stem "How much of the time during the **past 4 weeks** …" at the top of the page while the rest of each item appears in the grid below it. In Figure 2, the same item

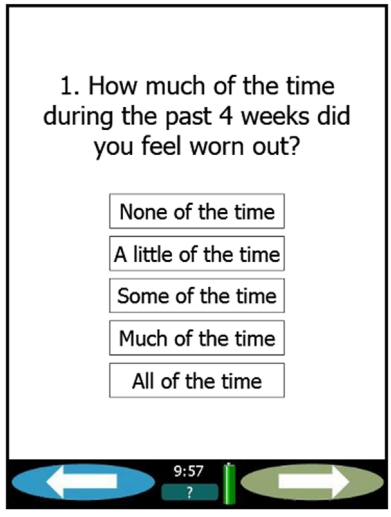

Fig. 2 – Example of item 1 postmigration in handheld device format.

is presented in an ePRO format and, in this case, item 1 now reads "How much of the time during the past 4 weeks did you feel worn out?" because the item and the stem have been joined together. In addition, because of screen space constraints, response options must be displayed vertically in the ePRO format (see Fig. 2) instead of horizontally as in the original presentation (see Fig. 1). Nevertheless, it may be possible in future ePRO implementations to utilize a landscape format in order to take advantage of a wider screen to enable horizontal presentation of response options, such as NRS and VAS. Figure 2 also illustrates a case in which the recall period "past 4 weeks" is not bolded on the handheld device format although it was bolded on paper because some platforms are not able to render this type of formatting in the electronic version.

Migration to a tablet device

Because of the larger screen area available, a decision needs to be made in consultation with the instrument developer regarding whether to present one item per screen or to present multiple items per screen on the tablet. With larger size tablets it may be possible to display the entire page as it appeared on the paper version, with radio buttons or checkboxes for responding. A grid format can be retained on the tablet, although it may not be possible to retain all items on the screen as on paper.

It is important to remember that a faithful migration does not necessarily mean that the electronic version must look exactly the same as the paper version. It means that the item and response text from the paper version have been retained, the instructions may need to be modified to reflect the new mode, and the migration does not affect the data the subject enters. Presenting multiple items per screen may save time with a very long questionnaire, but it also runs the risk of missing data or confusion. It may not be clear to the subject that multiple responses need to be selected on the same screen. Moreover, if functionality is not programmed to prevent skipping questions or advancing to the next screen without completing all questions, the risk of missing data is increased.

However, a tablet presentation allows for larger fonts and more space to display text, so there are fewer concerns over fitting the instructions, questions, and responses on the same screen. It should be noted, however, that although space may be available for multiple items per screen on a tablet, a single item per screen approach can provide consistency across multiple screen-based migrations of the instrument. Consistent presentation, that is, less variation in the presentation of an instrument's items across different size screen-based devices, is optimal and a recommendation of this task force, especially if more than one data collection mode is being used in a clinical trial.

Migration to a Web-based format

Although Web-based instruments appear to be device independent, there are still constraints contingent on the type of device used to access the Web interface. Because of the wide range of browsers, devices, and screen sizes, a decision needs to be made whether to allow certain types of browsers, such as those for mobile phone devices, to access the instrument. The screen design for a Web-based instrument is optimized for a specific browser and operating system or multiple browsers and operating systems, and therefore an instrument intended to be viewed on a desktop or laptop would be more similar to a tablet design due to the assumption of more screen space available. The screen design for a mobile device, however, would be closer to a handheld implementation and would require one item per screen formatting. Bring Your Own Device (BYOD) implementations [32] could be Web-based or developed as an "app" that resides on the subject's smartphone [13], and the migration implications would depend on which of these options was pursued. If a BYOD study was implemented in a Web-based format, then the same issues regarding optimization for a given browser, operating system, and screen size would apply.

Because of the wide range of screen sizes and formats with Web-based instruments, there may be greater risks of differences in the interpretation of scales. For example, the VAS is scored by measuring the point selected on a line. Line length will vary greatly by screen size and browser. This could lead to different responses in relation to the overall length of the line. As stated previously, it is best to use the single item per screen format because it provides the potential for less variation in presentation across different screen sizes.

Migration to an IVR system

For IVR systems, subjects respond to recorded scripts by using, primarily, zero through nine on the telephone keypad. For the most part, migration considerations for IVR have to do with the manner in which the item text (stem) or response options are formatted. For example, if the item is in the form of a statement, it may make more sense to rephrase it in the form of a question: "I feel tired" becomes "Did you feel tired?"

With regard to response options, if a verbal rating scale is used, responses must be associated with numeric entries and these must be incorporated into the IVR script for the instrument. In the case of an NRS that has each end of the scale anchored by a descriptor (e.g., 0 = "None" and 10 = "Worst imaginable"), but no descriptors in between, the script needs to describe that response context clearly. For example, such items are often worded as follows: Use a scale from zero to ten, where zero means none and ten means worst imaginable.

In addition, a traditional VAS, with verbal anchors on each end and no demarcations or descriptors at interim points, cannot be effectively operationalized on an IVR system without changing it to an NRS. On a technical level, if responses require pressing two or more numbers on the keypad (e.g., 10), the script should confirm with the subject the intended response because one of the numbers may not have been recorded by the system (e.g., only "1" was recorded rather than the intended "10" because of insufficient pressure on the keypad).

Usability versus Feasibility

When performing a migration from one data collection mode to another, establishing the subject's ability to use the new mode, or *usability testing*, is an important component of the migration. Coons et al. [29, p. 423] stated:

> Usability testing examines whether respondents from the target population are able to use the software and the device appropriately. This process includes formal documentation of respondents' ability to navigate the electronic platform, follow instructions, and answer questions. The overall goal is to demonstrate that respondents can complete the computerized assessment as intended.

Usability testing is an indication of the subject's ability to navigate or use a particular data collection system. Because it is focused on the respondent's ability to use the system, it may be conducted at an investigative site in a controlled environment with observation of the subject.

Although usability testing is always recommended to establish subjects' (or end users') ability to use the system, *feasibility testing*, or the evaluation of the system within a specific study design, may be necessary only in certain circumstances. The distinction between these two types of testing is best characterized as follows: usability testing assesses whether the data

collection mode can work under general conditions, whereas feasibility testing assesses whether it will work in the context of a specific study design or a specific instrument.

The need for feasibility testing will be driven by the novelty of the study design in which the PRO data collection system is to be implemented. For example, if the system is to be implemented for site-based PRO data collection in a standard study design that has been previously implemented in numerous trials, there may be no need for additional feasibility testing. As a counter example, if the system is to be implemented in a novel study design, in which field-based data are being collected in a unique way (e.g., multiple times per day) for a given patient population, then feasibility testing will ensure that the PRO data collection system actually works with the patients in the study design and using the new instrument. Thus, the evaluation of whether or not to conduct feasibility testing, in addition to usability testing, will be on a case-by-case basis and driven by the novelty of the study design and the instrument.

If feasibility testing is deemed necessary, the testing plan should include recruiting subjects similar to those who will participate in the clinical trials; subjects following the study procedures as required by the study design for a reasonable period of time (e.g., using the diary for 7 days); and then performing debriefing interviews with the subjects to assess their compliance with the study procedures (e.g., whether they completed the diary every day, as requested) as well as to assess usability. The debriefing of the subjects is best facilitated by review of actual compliance data captured during the study and likely reported on a portal of some kind. As with cognitive interviews for migration equivalence, a sample size of 10 to 15 subjects should be sufficient.

It is important to note that neither usability testing nor feasibility testing as described above is the same as another process called *user acceptance testing* (UAT). (UAT is one aspect of an extensive system/software validation process that is far beyond the scope of this article. Another ISPOR PRO Task Force report, Validation of Electronic Systems to Capture Patient Reported Outcomes (PRO) Data—Recommendations for Clinical Trial Teams: A report of the ISPOR ePRO Systems Validation Task Force [33], addresses this topic.) According to Coons et al. [29], "the purpose of UAT is to determine whether the software complies with the written system specification or user requirements document." It is not intended to determine whether respondents like or can use the system. UAT does not include clinical study subjects. We recommend that usability testing and, if necessary, feasibility testing occur in addition to UAT following a migration to a second data collection mode.

Equivalence

Any migration process involves some type of modification(s) to implement the instrument in the new mode. The goal of the faithful migration is that subjects interpret and respond to the questions/items on the PRO instrument the same way regardless of the data collection mode. Once a migration has occurred, it is necessary to determine whether this goal was achieved through an evaluation of the measurement equivalence between the original and migrated modes. The previous ISPOR Task Force report [29] focused on the degree of modification as the key factor in determining the level of evidence needed to establish equivalence. In this section, we build upon the work of that previous task force and recommend additional considerations for determining the level of evidence needed to establish equivalence. We also delineate the types of equivalence testing and the typical procedures to execute such work.

Need to Establish Equivalence

One consideration is whether measurement equivalence needs to be established between the original and new data collection modes. In the context of instruments that will be used in registration trials for submission to the FDA, measurement equivalence should be established and documented by the study sponsor if the data are to be used to support labeling for a medical product. If there are sufficiently rigorous published data to support that equivalence, then further equivalence studies are not needed. A decision tree is shown in Figure 3. Developer requirements may supersede these recommendations if a certain type of equivalence study or level of evidence is a condition for use of the instrument. It is worth noting that from a scientific perspective, we believe that it is always necessary to have confidence, through evidence, that measurement equivalence exists because this has a direct impact on interpreting any results from a migrated instrument (represented by always following the right-hand side of Figure 3).

Levels of Equivalence Evaluation

Following the faithful migration and the determination that equivalence needs to be established, the appropriate level of

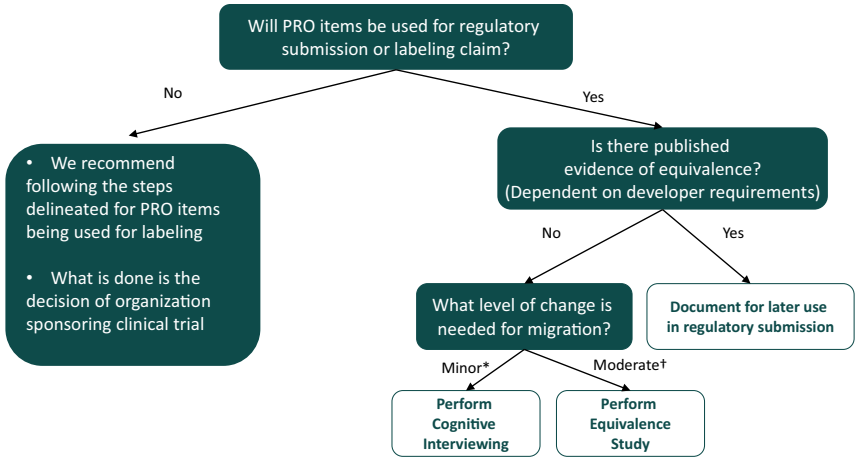

Fig. 3 – Decision tree regarding the need to establish measurement equivalence. PRO, patient-reported outcome. *Minor, changes to the instrument are not likely to have changed the interpretation or responses. †Moderate, changes to the instrument may have changed the interpretation or responses.

equivalence evidence needs to be identified. The level of equivalence evidence is dependent on the extent that changes or modifications are likely to have had an effect on the subjects' interpretation and responses to the items in the instrument.

Table 1 summarizes the levels of modification that might occur during faithful migration and is an adapted version of the one presented by Coons et al. [29]. In the course of performing the migration, two types of changes may occur and need to be considered when equivalence is evaluated: format and procedural. *Format changes* refer to differences between the modes in terms of format, including how the items and responses are presented to the subject. For example, formatting modifications include adapting instructions from a paper to an electronic mode, such as changing "circle" to "select." *Procedural changes* refer to the different ways modes are actually implemented in studies and include aspects such as edit or validation checks, introducing a jump or skip sequence so that subjects do not see questions that are not relevant to them, completion windows, and compliance with protocol requirements such as when to complete data collection. In general, procedural changes between modes may have a greater effect on how the subject responds to questions because the electronic modes can limit possible responses to those that are within an appropriate range, prevent unintentionally skipped questions, and enforce completion windows; none of these is/was possible with paper.

Table 1 – Levels of modification and equivalence.

Level of modification	Rationale	Examples	Level of evidence
Minor	The changes to the instrument are *not likely* to have changed the interpretation or responses.	Format: 1) Nonsubstantive changes in instructions (e.g., from circling the response to touching the response on a screen). 2) Minor changes in format (e.g., one item per screen rather than multiple items on a page). Procedural: 1) Implementation of tablet at the site with differences in edit checks, validation rules, and branching logic.	Cognitive interviewing Usability testing
Moderate	The changes to the instrument *may* have changed the interpretation or responses.	Format: 1) Changes in item wording or more significant changes in presentation that might alter interpretability (e.g., spreading an item over two or more screens because of space constraints, changing the structure of the response options). 2) Change in the mode of administration involving different cognitive processes (e.g., paper [visual] to IVR [aural]). 3) Change in the mode of administration to Web-based administration (e.g., variance between screen sizes too great to be considered minor modification). Procedural: 1) Migration of paper diary to electronic platform with differences in completion windows, compliance with planned assessment schedule. 2) Differences in the ways that subjects are alerted to complete instruments (e.g., alarms on a handheld device always available vs. e-mail reminders for Web that require logging into e-mail are not as proximal to the actual reminder time vs. no reminders at all on paper, so compliance could differ).	Equivalence study Usability testing

IVR, interactive voice response.

The breadth and extent of the modifications made during the migration process, some of which are necessary and others that facilitate easier administration of the instrument, will vary in terms of their effect on influencing the subject's interpretation and responses. Table 1 incorporates examples of minor and moderate levels of procedural as well as format modifications to illustrate how both types of changes can be evaluated in terms of levels of equivalence. Minor modifications have a low likelihood of affecting interpretation and response, and therefore cognitive interviews and usability testing are sufficient to confirm the equivalence between the modes. Moderate modifications introduce the possibility of affecting interpretation and response, and therefore it is recommended that a quantitative equivalence study along with usability testing be conducted to evaluate equivalence between the modes.

It is critical to determine the level of equivalence needed as part of the equivalence study planning process. Table 1 is based on Coons et al. [29], and additional detail has been added to illustrate what constitutes minor and moderate modifications. In cases in which modifications fall into more than one level, the recommendations associated with the higher level of modification and, therefore, evidence should be followed. It must be noted that if substantial changes to the item content (e.g., recall period) and/or response options are needed to enable the migration of the instrument to a new mode of data collection, the instrument is considered a new instrument and full psychometric evaluation would be necessary.

Types of Measurement Equivalence Studies

Two major types of studies may be conducted to evaluate measurement equivalence between modes:

1. Qualitative studies involve cognitive interviews that provide qualitative data to evaluate equivalence between modes. These studies have previously been associated with a minor degree of modification between modes.
2. Quantitative studies are intended to evaluate statistical equivalence of responses, involve much larger sample sizes, and focus only on the statistical comparison of responses to both modes of the PRO measure.

Common qualitative study designs

Qualitative studies involve small samples of 10 to 15 participants who are from the target population of the confirmatory clinical trial, usually phase 3. Qualitative study designs are used to evaluate the effect of format changes between the original mode and the migrated mode to ensure that the subject's interpretation of the items on the migrated mode is comparable to the original.

Cognitive interviews during the instrument migration process

It bears pointing out that cognitive interviews conducted during the instrument migration process serve a different purpose than do cognitive interviews conducted during the instrument development process. In general, cognitive interviewing techniques are used to study the way in which subjects "understand, mentally process, and respond to materials presented to them" [34]. Of interest here are the cognitive interviews conducted subsequent to instrument migration, which are aimed at determining whether subjects are interpreting and responding to the items the same way on the new mode as they would on the mode from which the instrument was migrated. See the following text for different cognitive interviewing approaches. In contrast, cognitive interviewing during instrument development is primarily aimed at supporting the instrument's content validity by determining whether subjects are interpreting items and using response scales as intended. In the current context, cognitive interviews are not intended to revisit the content validity of the original instrument.

One approach to conducting cognitive interviews involves having subjects complete the instrument on the original mode and the new mode of data collection and determining whether there are items for which the responses differ between the two modes. A distraction task can be included between the completion of the modes to reduce potential memory/carryover effects yet allow a short interval between administrations to reduce subject burden. The interview then focuses on those items individually to determine whether the different responses were random (i.e., "I could go either way") or systematic due to a difference in the meaning or interpretation of the item by the subject on the alternative modes.

If the latter is the case with a substantial number of subjects in the cognitive interview sample, the changes made in migrating those items to the new mode need to be revisited to determine whether a successful migration of those items is possible. It should be noted that, other than using the responses on the two modes to identify where differences exist, this approach is not quantitative; the responses are not used for any descriptive or inferential statistical analyses. Furthermore, if such discrepancies are significant enough to warrant change to the newly developed mode, then cognitive interviewing must be replicated to ensure that the change resolved the discrepancy. If there are still discrepancies in subject qualitative reporting due to format differences, a quantitative equivalence study should be considered (see below).

A second approach involves having subjects complete the instrument on the new mode and asking them how they interpret what each item is asking them. This can be accomplished by asking the subject to repeat the question being asked in his or her own words (i.e., paraphrasing) or through a think-aloud task that involves the subject talking through how he or she arrives at the response [34]. The subject's interpretation of the item is then compared with the item definition or concept elaboration document prepared by the instrument developer to determine whether there is concordance. This approach more closely parallels cognitive interviewing during the instrument development process. It assumes that documentation of the intended meaning/interpretation of the items is available. If the instrument had been translated for use in other languages/cultures, such documentation should exist because it is essential for linguistic validation. If it does not exist, it should be possible to construct it in conjunction with the instrument's developer.

A third approach is to ask subjects only about instructions and/or items that were modified during the migration process to reduce subject burden and interview length. This enables a more focused investigation of the potential effect of those changes and a potentially shorter interview. Subjects are asked to read both versions of the instructions or items on the two modes and identify any perceived differences in the self-report task or in the interpretation/meaning of modified items. If most or all of an instrument's items required modification during the migration process, however, then this approach does not necessarily decrease the amount of time required to conduct the cognitive interview because all such items would still require debriefing.

At the present time there is no consensus regarding the optimal approach to cognitive interviewing during the migration process. A combination or hybrid of two or more of the above is a viable option if it makes sense for a particular study.

Common quantitative study designs

Quantitative equivalence studies are recommended for moderate modifications between the modes (see Table 1) when migrating and for mixing modes that involve visual versus auditory use (IVR), use

of Web at subjects' homes, and for paper versus electronic diary (field-based assessments) studies. All these scenarios present greater risks for differences in response between modes and therefore a greater need to demonstrate that they provide sufficiently equivalent results.

Within quantitative equivalence study types, Coons et al. [29] mention randomized crossover and randomized parallel groups as the typical options for evaluating equivalence. Randomized crossover designs have become the preferred study design for migration equivalence studies because subjects serve as their own controls and therefore the sample size is significantly reduced. Within the randomized crossover approach, study designs may be either single visit when evaluating whether the migration changed interpretation or multivisit when evaluating whether migration changed interpretation and how items were completed in the context of implementation. The multivisit study design is most useful for evaluating field-based assessments that are intended to be completed on a daily basis over a period of time and scores are typically averaged.

Figure 4 depicts the most common study designs for quantitative equivalence studies. The single visit study design is appropriate for site-based or field-based assessments and involves a randomized crossover in which each study subject completes both modes of data capture but is randomized to the order of completion (i.e., randomized to which mode is completed first to control for order effects).

The multivisit field evaluation study design is appropriate for diary (field-based) instruments if there is a need to establish that the two modes of data collection are equivalent in the context of a simulated study design. The multivisit study involves a randomized crossover in which each subject is randomized to order and completes the first mode for 1 to 2 weeks and then crosses over to complete the second mode for the same length of time. This approach is recommended in cases in which the two modes are intended to be mixed in future studies.

Because of the longer study duration, it is important to also assess whether the subject's condition has changed to ensure that the comparisons made are only within those subjects who have not changed during the course of the study. This approach allows for procedural differences between the modes to be tested in a setting similar to what the subject would experience in the clinical trial. It has an increased risk of demonstrating a lack of equivalence, however, because of the potential for larger response differences between the two modes due to the manner in which the subjects are completing data entry.

For example, if a protocol has subjects entering data within a specific time window per day, data entry with an electronic data collection device may be confined to that time window whereas with paper-based data collection, data may be entered by the patients at any time. This study design functionally includes the concept of feasibility, examining performance of the system in the context of a specific study design, as discussed above. Therefore, the situations in which a multivisit equivalence study design is incorporated should be carefully considered.

Single visit studies answer a different question from multivisit study designs. Single visit studies focus on equivalence in interpretation at a point in time, which is sufficient when moving away from the paper-based data collection, whether at the site or in the field. Multivisit studies are needed to address equivalence between modes in a field-based context of a specific study design, and are needed if intending to mix modes in the future. More specifically, if one intends on mixing paper and electronic diaries in a trial, the multivisit feasibility study is needed to establish the equivalence of these two modes of PRO data collection in a real-life setting, given the procedural differences between the two modes. Again, this will also functionally accomplish the goal of feasibility testing of the electronic data collection mode. See Table 2 for a comparison of these two study design approaches for quantitative studies.

In either case, the main statistical method used and recommended in the evaluation of the equivalence between responses is the intraclass correlation coefficient (ICC) as discussed in Gwaltney et al. [17], Coons et al. [29], and McEntegart [35].

In cases in which little evidence of the test-retest reliability of the original version of the instrument is available, it may be helpful to conduct a "double-cross" study in which each subject crosses between modes and then back to the original mode so that test-retest reliability can be obtained and compared both within and between modes. See Appendix C in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.06.005 for details on the "double-cross" study design.

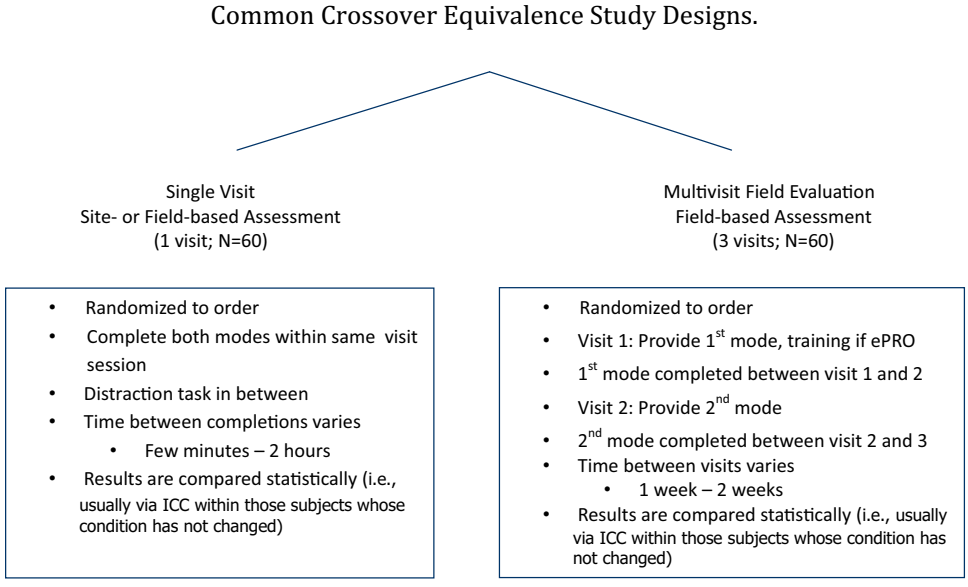

Fig. 4 – Common crossover equivalence study designs. ePRO, electronic patient-reported outcome. ICC, intraclass correlation coefficients.

Table 2 – Two common study design types and some considerations for each.

Instrument type	Study design type	Pros	Cons	Limitations
PRO instruments completed at site; Field-based assessments in which mixing is not intended	Single visit—randomized crossover	Statistical equivalence level between modes can be established	Assesses format differences but not procedural differences	Comparison with original mode test-retest reliability may be limited; does not reflect performance of paper diary in clinical trial setting
Field-based assessments, especially frequent or episodic assessments per day, where mixing is intended although not recommended	Multivisit field evaluation—randomized crossover	Statistical equivalence level between modes can be established; real-world setting for field-based assessment	Studies difficult to operationalize because target concepts are variable, need to control for change; high likelihood that equivalence won't be found	Comparison with original mode test-retest reliability may be limited

PRO, patient-reported outcome.

When planning subject recruitment for qualitative and/or quantitative equivalence studies, it is important to consider potential overlap with recruitment for clinical trials with the same patient population, especially in rare diseases for which the population for trials is limited. It is acceptable to recruit subjects for equivalence studies who may then go on to participate in a clinical trial in which the modes tested in the equivalence studies are to be used, in order not to reduce the pool of potential clinical trial participants. Unless it is absolutely necessary, it is not recommended to recruit participants for equivalence studies who have already participated in a clinical trial or validation study using one of the modes in question because they have already experienced the mode and PRO instrument being studied and may have a biased response during the equivalence study.

Qualitative study designs are acceptable for demonstrating measurement equivalence for minor modifications and for migrations in which the original and alternative data collection mode are not intended to be combined in clinical trials. These studies do not statistically test measurement equivalence for mixed modes and are insufficient for mixed paper and electronic field-based assessments (e.g., daily diary) to be used within the same trial. Within quantitative study designs, if a field-based assessment is tested in a clinic-based single visit design, it does not reflect the actual trial setting and is unlikely to assess the true performance of the instrument.

It is also critical for field-based assessment studies that the subject population is stable and unchanging to limit true change in response in equivalence studies, but clinical trial use assumes that subjects will change over time because of the treatment. Therefore, it may be impossible to distinguish what is driving change in scores when mixed paper and electronic field-based assessments are used in a treatment setting. The result of the equivalence studies may be to conclude that potential differences between paper and electronic field-based assessments are too great to allow mixing modes within a clinical trial, and in these cases the default should be the electronic data collection mode only.

Because the term *migration* in and of itself merely refers to the transfer of an instrument from one mode or format to another, it carries no implication of what will be done with either mode in the future. In many cases, the migration results in a new mode that will replace the original mode in future studies, while mixing involves using both old and new modes within or between studies and then pooling the data from different modes for analysis. Therefore, when migrating permanently it is only necessary to demonstrate equivalence for prospective use, whereas when migration results in mixing modes, it is necessary to demonstrate equivalence for concurrent use. In the former case, a qualitative study may suffice, whereas in the latter case it is necessary to conduct a quantitative equivalence study. Therefore, it is strongly recommended that the potential for mixed modes be considered at the start of the equivalence study planning process so that the appropriate approach to evaluating equivalence, qualitative or quantitative, will be used in the most timely and efficient way.

Mixing

Although mixed modes can and do occur in all research settings, the primary focus of this article is on clinical trials in which the PRO end points are intended to support labeling claims. The discussion thus far regarding mixing modes has focused on mixing within a given trial. The general recommendation is to avoid, where possible, such mixing because of the increase in measurement error associated with introducing any variable into a study.

We note that *mixing*, as used in this report, refers to the administration of the same instrument via different data collection modes in a single clinical trial; it does not refer to the administration of different instruments via different data collection modes in a single clinical trial. The latter does not pose a threat to measurement error as discussed in this report.

Our recommendations thus far for mixing have focused on determining the need to establish measurement equivalence when mixing occurs. It could be argued that randomization of subjects into groups is sufficient to account for mixing. As long as the pattern of mixing modes is the same in treatment and control groups, any potential measurement error introduced by the mixed modes will be comparable across the two groups. However, even the balanced introduction of measurement error across treatment arms has the potential to put the trial at risk of not showing a treatment effect if the signal to noise ratio is decreased. Any change during the trial (after randomization) that leads to different data capture mode patterns across the treatment and control patients (or within treatment or

control patients) has the potential to differentially introduce measurement error.

There are a number of ways in which mixing of modes can occur in the development of medical products, including mixing

1. between product development programs,
2. between clinical trials within a program, and
3. within a single clinical trial, such as
   a. countries within a trial,
   b. sites within a country,
   c. subjects within a site,
   d. within a subject,
   e. and time points within a trial (e.g., start with one mode and change to another mode).

We now turn to a discussion of the various ways in which modes might be mixed.

Mixed modes occurring between medical product development programs or clinical trials within a program are often the result of evolving technology. New and better methods of PRO data collection emerge or regulatory requirements change and modes may change during the product development program. The implementation of mixed modes in these situations can be carefully planned and executed with supporting studies to demonstrate the equivalence of modes.

In this manner, if the old or original mode is abandoned, a demonstration of measurement equivalence supports consistency of interpretation of the data between trials, and no additional activities are needed in new trials because only a single mode of PRO data collection is being incorporated. In some development programs, there may not be a need to compare current trials to previous ones with respect to the PRO data. If that is the case, then there is no need to establish measurement equivalence between the new mode and the one used in previous trials. The need for measurement equivalence will be driven by the need to compare data across trials within the program.

The remaining types of mixing are within a given clinical trial. The first that we will consider is mixing modes across countries within a trial. When conducting multicountry studies, not all countries may have access to the technology being implemented for PRO data collection (e.g., Internet access for Web-based version). In such cases, specific countries within multicountry studies may need to collect PRO data using one mode while the remaining countries use another. Measurement equivalence should be demonstrated across mode. Historically, researchers considered mixing technology-based solutions with paper-based solutions if a specific country was not able to support the selected technology.

With careful planning, however, the modes that are mixed may have only minor differences (e.g., both are screen-based systems), and thus, a lower likelihood of introducing measurement error resulting from measurement inequivalence. For example, if a Web-based PRO data collection system is the default for the trial, perhaps a handheld device or tablet can be used for countries that do not have Internet access that can support the Web-based system.

Within a country, modes can also be mixed between participating clinical sites within a trial, again, because of access to the specific technology, or possibly the site's ability. If the issue is site ability, then the potential mixing of electronic solutions becomes more challenging. The likely case is that the investigative site does not believe that it can implement the technology-based solution. It may be possible to have more similar technology solutions that a site can implement, which, again, will minimize the potential for introducing measurement error. Alternatively, the sponsor may choose to mix paper and electronic solutions in the trial. If the PRO data will be field-based assessment, our general recommendation still prevails; it is unlikely that equivalence can be established. In such cases, it may be prudent for the sponsor to consider other options such as not including the specific site or region in question. (This also applies to the country case described above.) In contexts in which the subject sample is extremely difficult to recruit, such as for rare diseases, the sponsor may be faced with a significant dilemma between mixing modes and increasing measurement error, versus obtaining the subjects for the trials. In such cases, the sponsor will need to make this decision on the basis of the specific issues facing them.

Another common situation of mixing between sites occurs when site and subject recruitment proceed faster than anticipated and the electronic system is not yet available at the time the subjects are enrolling in the study at the first few sites. This is particularly an issue in indications that are seasonal in nature such as seasonal allergies, chronic obstructive pulmonary disease exacerbations, and influenza, for which patient recruitment cannot be delayed until the electronic system is ready for deployment. Despite the sponsor's best intentions to use an electronic data collection mode, such situations may require an interim solution, often paper, until the electronic system is validated and ready for launch so that subjects can be enrolled to meet clinical trial timelines.

Subjects within a given investigative site may have the need for various modes of PRO data collection because of subject ability, preference, health state, or site preference for a given subject. Such decisions to mix subjects within a given site should have been anticipated and planned for in a manner similar to the above cases of mixing across countries or sites. If mixing of modes within a site is anticipated, then appropriate upfront equivalence should be established.

The additional challenge, however, is deciding that a subject needs one mode or another. We can anticipate that in these cases, the request will be for a familiar mode of PRO data collection, likely paper. If a potential subject has never previously used an electronic method of data collection, then his or her initial preference may be to select paper if that is the other option presented in the trial. Such subject preference would not be based on actual ability to utilize the electronic solution, but a subject's impression or belief. We recommend that rather than letting the subjects decide, a more objective method of evaluating individual subject's ability be used if a sponsor wants to provide options to subjects. Such an evaluation can take the form of having the subject attempt to use the electronic method at the investigative site. If this evaluation demonstrates that the subject cannot use the primary electronic solution, possibly because of subject ability or state of health, then the investigative site should establish that the subject can use the alternative mode. It should be emphasized to the investigative site that this evaluation is mandatory so that the site does not attempt to use its own, idiosyncratic evaluation of subject ability/competence to make this decision. We recommend that mixing modes of PRO data collection within an investigative site be used rarely and approached cautiously.

In the final situation for consideration, subjects may begin a study using one PRO data collection mode and finish with another. Some trials have built in a paper backup solution for situations in which the technology fails, given that one potential issue for any electronic PRO data collection system is failure of the technology. Thus, some subjects may begin in one mode but switch to another as a backup in cases of device loss or failure, or inability to access the electronic version. In other cases, the subject may begin on paper before the electronic mode is available because of programming and validation delays, and then switch to the electronic mode later. In such situations, it will be important to note where such a switch took place within the duration of the trial.

Table 3 – Mixing data collection modes and risk of not having equivalence.

Level of mixing	Risk to equivalence	Comments
Between product development programs	Varies	If there is no need to compare or pool current results with previous products developed, then risk is low. If the new product is in the same therapeutic area, the risk may become higher. Analytical techniques can be used to evaluate any error introduced because of mixing; see the *Post-trial* section.
Clinical trials within a program	Varies	The risk will vary on the basis of the stage of product development and the need for trial comparability or pooling. Analytical techniques can be used to evaluate any error introduced because of mixing; see the *Post-trial* section.
Countries within a clinical trial	High	If data are to be pooled across countries, as with most trials, comparability is very important and differences between countries should be evaluated.
Sites within a trial	High	Difficult to establish between-site comparability of data if mixing occurs; therefore, it is discouraged.
Subjects within a site	Very high	It is very difficult to assess a site's performance if the data are collected with different modes by subjects at the same site. It may be very difficult to determine whether changes are due to response to treatment or due to difference in mode. **This is not recommended.**
Within a subject	Extremely high	Can potentially compromise usability of subject data. Difficult to demonstrate that mixing did not have impact. Strongly discourage this level of mixing except in extreme cases in which data would otherwise be missing. **This is not recommended.**

Any one of these types of mixing can yield differences in the data and introduce measurement error into the trial or clinical program results. We therefore recommend that the need for data comparability and the impact of introducing measurement error be assessed for each situation. Our recommendation is that if data are to be compared or pooled at any level, quantitative evidence of equivalence is necessary. Table 3 provides a presentation of the risk of not having equivalence at each level at which mixing may occur.

Table 3 suggests situations that vary in risk when using mixed modes for measurement of PRO end point(s) between and within a clinical trial. This could be the difference between success and failure for the trial if the PRO is the primary end point. As mentioned in the Introduction, not having measurement equivalence between the modes could increase measurement error, in turn attenuating the ability to identify a treatment effect within a given trial. Such measurement error could then result in a nonsignificant difference in the primary efficacy PRO end point for a new medical product.

Types of Modes Being Mixed

Once it is determined that mixing will be done either between or within a trial, there are considerations for which types of modes to mix. We now turn to issues to consider with mixing various types of modes of PRO data collection across the situations just described. This includes mixing paper and electronic modes and different electronic modes of data collection.

Mixing paper and electronic modes

Mixing paper and electronic modes is the most risky combination because of the differences in how a subject interacts with paper, having little to no restriction on how he or she responds to questions, in comparison with electronic modes, which verify responses via edit checks and restrict the subject in how he or she may respond to questions (e.g., allowing only one response per item or not allowing items to be skipped before proceeding). Another caveat to mixing paper and electronic modes is the cost of building and validating two separate databases and a separate data entry system for paper, which requires double data entry and extensive querying and data cleaning. The logistics of paper data entry in the context of an electronic clinical trial raise additional questions regarding who is responsible for the paper data entry and having it completed in a timely manner so that trial timelines are not delayed. Our general recommendation is to avoid mixing paper and electronic modes of data collection to the extent possible. There is less risk in mixing site-based instruments because they are completed under supervision and corrections to invalid responses on paper can be made, and if equivalence between these modes has been previously demonstrated, but the cost and logistics caveats remain.

Empirical evidence is emerging to demonstrate moderate to strong correlations between paper and Web-based data collection for site-based assessments [21–24]. However, studies with the WOMAC [28] and the Bath AS scales and Quebec scale [27], all typically administered as site-based assessments have shown a tendency toward lower scores on the computer version than on the paper version, leading Bent et al. to recommend that "the small tendency for the computer format to score lower than the paper format may indicate that the same format should be used in any subsequent retesting with the same outcome measure" [27]. We strongly discourage the mixing of paper and electronic field-based assessments because of the significant potential equivalence issues, the significant procedural change between these two modes, and the likelihood that they will not generate equivalent responses.

The FDA clearly discourages field-based PRO data collection using paper because of the inability to know when the data are entered. Specifically, the PRO Guidance [1] states:

> If a patient diary or some other form of unsupervised data entry is used, we plan to review the clinical trial protocol to determine what steps are taken to ensure that patients make entries according to the clinical trial design and not, for example, just before a clinic visit when their reports will be collected.

This quote specifically addresses not only the ill-advised use of paper as the single data collection mode but also underscores the impracticality of mixing paper and electronic diaries in a clinical trial setting.

Mixing electronic modes

Mixing visual modes, such as tablet with the Web or a smartphone device, is less risky because it is potentially easier to demonstrate equivalence between these modes and implement them in a similar fashion so that differences in format are minimized. The use of the Web without restrictions on screen size and resolution, however, is potentially risky because it is not possible to control all elements of the visual presentation across browsers, operating systems, and device types to ensure that all subjects see the questions and responses the same way. Bring Your Own Device implementations are another example of mixing within electronic modes, and similar issues apply in the case of mobile Web implementations because of the range of screen sizes and device types that can be used to access the instrument. We recommend caution when using Web-based data collection for this reason and consider it a mixed modes situation because of the degree of differences.

Mixing visual and auditory modes, such as Web and IVR, requires that quantitative equivalence be demonstrated to ensure that the moderate difference between modes does not affect interpretation and response. Some studies have demonstrated equivalence between Web and IVR modes [23,24]. There may also be implementation challenges with such disparate modes that need to be considered.

From a cost perspective, mixing Web and IVR is often cost-effective because the data are stored in the same database on the backend, so the incremental cost of adding a second mode is low; most IVR systems have a Web component built in. The cost of mixing other electronic modes depends on whether two entirely separate systems and databases need to be built or whether some kind of efficiency on the backend is available.

Operational and Statistical Considerations for Mixing Modes

At this point, once a decision to incorporate mixed modes within a trial, and which modes will be mixed, has been made, there are a number of operational and statistical issues to be considered as implementation occurs.

Pretrial preparation

As discussed many times in this report, it is important in most situations to first evaluate equivalence between the modes. It is strongly recommended that the potential for mixed modes be considered at the start of the equivalence study planning process so that the appropriate approach to evaluating equivalence, qualitative (for minor and/or permanent migration) or quantitative (moderate and/or mixed modes in the future), will be used in the most timely and efficient way. It is also critical to assess the risks of certain types of mixing as described above. Assuming that measurement equivalence between the modes has been established, the results of this evaluation should be taken into account when determining the sample size for the study. This can be done by working with the appropriate biostatistician to adjust the presumed measurement error in the sample size calculation. The specifics of this computation are beyond the scope of this report.

In addition, pretrial planning should include the issues raised above in the discussion of mixing within a trial. That is, appropriate training for both modes will need to be conducted. Perhaps most important, the criteria for determining which countries, regions, sites, or subjects are permitted to mix will need to be established, documented, and then clearly conveyed to the investigative sites.

Trial implementation

If mixing was preplanned, then one of the key implementation challenges will be to manage where and when each mode is used. If the plan was for mixing across countries, regions, or sites, the challenges will be fewer than cases in which mixing occurs within a site or within a patient because the data collection mode will be same within the country, region, or site. As mentioned above, the key issue will be to ensure that the investigative site follows the sponsor's procedures for mixing rather than using idiosyncratic criteria.

Perhaps the most challenging situation will be the case in which mixing was not planned, an electronic PRO data collection mode fails, and the sponsor defaults to a paper-based method. Our recommendation is that the sponsor should always have a contingency plan in case of technology failure, which involves procedures to replace the same mode quickly so that downtime and missing data are reduced. In the absence of such planning, and some determination of the potential impact of mixing in the case of electronic system failure, ad hoc implementation of a backup may compromise the study data. Because of major risks in this type of ad hoc mixing in studies involving field-based data collection, options other than paper should be considered as a backup in case there is device loss or failure, mainly recovery or replacement of the same mode. The primary issue is the potential for missing data versus the introduction of measurement error through mixed modes of data collection. The sponsor will need to consider a backup solution, and what solution most appropriately balances those two considerations. The sponsor may decide that a backup solution that minimizes missing data takes priority regardless of the nature of the backup. In addition, a low level of mixing of modes (e.g., $<10\%$) may not have an impact on the overall result, but sensitivity analyses are recommended to verify whether it did or not, especially if equivalence has not been shown *a priori*. Finally, whether mixing is planned or happens ad hoc, it will be necessary to develop the statistical analysis plan (SAP) to address the analysis of mixed modes *a priori* to evaluate whether the treatment effect differs by mode.

Post-trial

In cases in which there were mixed modes in a trial, the best-case scenario is that the mixing was planned and controlled and that there is previous evidence of measurement equivalence. In this case, the analytic methods should have been specified in the SAP and these methods should be applied. Because equivalence has been established, the primary analysis will likely include all data pooled irrespective of mode. Also, it will be prudent to conduct additional analyses to explore any effects of mixing. If there are sufficient data available per mode, it is recommended to compare summary statistics by mode and also include mode as a variable in the statistical model. Data permitting, it may also be appropriate to test for a treatment by mode interaction in the statistical model; such tests have low power but may suggest the need for further investigation. For instance, it may be helpful to conduct a sensitivity analysis to evaluate the impact of including or excluding the alternate mode data on the treatment effect. Any interpretation of a treatment by mode effect might be difficult but it is important to conduct the investigation to better understand the data and generalizability of the treatment effect. Gallo [36] gives a good summary of the considerations and analyses that apply to treatment by site interactions, and this topic has relevance to the situation of mixed modes. Other examples in which poolability would be tested include translations [37], countries, and/or regions.

In scenarios in which mode equivalence has not been established or the mixing was not planned and/or less controlled, it will be important to give careful consideration to how to amend the SAP. The type of mixing and volume of data on each mode will determine the types of analysis that are appropriate. Pooling with tests for treatment by mode interaction may be appropriate but even in the absence of a suggestion of any interaction, the results for any treatment effect from these analyses should be viewed as exploratory rather than confirmatory.

Conclusions

One of the most important developments in the field of PRO measurement in clinical trials has been the emergence of technologies that enable electronic collection of data. With the increasing variety of data collection modes, mixing these modes between and within clinical trials in a medical product development program is possible. Although it has the potential to add measurement error if not planned and implemented properly, mixing of PRO data collection modes within trials does occur and must be addressed pragmatically.

This task force report provides an overview of important issues to consider in the process of migrating between modes of data collection and also in using multiple modes of data collection in clinical trials. The key drivers of this report are to address the FDA concern regarding measurement equivalence and the potential impact of using multiple modes of data collection or administration on the treatment effect in the trial and the choice to reduce measurement error where possible when insufficient information about measurement equivalence is available.

It is important to conduct a faithful migration from the original mode to a new mode or modes to ensure that subjects interpret and respond to the questions/items on the PRO instrument the same way regardless of the data collection mode. The levels of evidence needed to establish equivalence have been expanded to address both format and procedural differences that occur between modes, and an overview of common study designs for both qualitative and quantitative equivalence studies has been presented.

In the absence of documented evidence of measurement equivalence, it is strongly recommended that a quantitative equivalence study be conducted prior to mixing modes in a trial to ensure that sufficient equivalence can be demonstrated to have confidence in pooling PRO data collected by the different modes. However, we also *strongly discourage the mixing of paper and electronic field-based instruments* and suggest that *mixing of only electronic modes be considered for clinical trials and only after equivalence has been established*.

If proceeding with mixing modes, it is important to implement data collection carefully in the trial itself in a planned manner at the country level or higher and minimize ad hoc mixing by sites or individual subjects. Finally, when mixing occurs it must be addressed in the SAP for the trial and the ability to pool the data must be evaluated in order to then evaluate treatment effects with mixed modes data. A successful mixed modes trial requires a faithful migration, measurement equivalence established between modes, and carefully planned implementation to minimize the risk of increased measurement error.

Acknowledgments

The members of the ISPOR PRO Task Force on Mixed Modes of PRO Data Collection were Sonya Eremenco (Chair), Ethan Basch, Antonia V. Bennett, Stephen Joel Coons, Karin Coyne, Cindy Gao, Zhanna Jumadilova, J. Jason Lundy, Damian McEntegart, Willie Muehlhausen, Jean Paty, Tara Symonds, Simrandeep K. Tiwana – Sidhu, Hwee Lin Wee, Kathleen Wyrwich, and Vladimir Zah. The steady and capable support of ISPOR and its staff, particularly Elizabeth Molsen, is genuinely appreciated. In addition, the contributions of Paul O'Donohoe and Jeff Sloan are gratefully acknowledged.

Supplemental Materials

Supplemental materials accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2014.06.005 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] US Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. 2009. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. [Accessed June 1, 2014].
[2] US Food and Drug Administration. Clinical Outcome Assessment Qualification Program. October 11, 2013 update. Available from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm. [Accessed June 1, 2014].
[3] Walton MK, Powers JH, Hobart J, et al. Clinical outcomes assessments (COAs): a conceptual foundation. ISPOR ClinRO Task Force report. Value Health. In press.
[4] US Food and Drug Administration. Measurement in clinical trials: review and qualification of clinical outcome assessments. In: Public Workshop; Silver Spring, MD; October 19, 2011. Available from: http://www.fda.gov/Drugs/NewsEvents/ucm276110.htm. [Accessed June 1, 2014].
[5] Hyland ME, Kenyon CA, Allen R, Howarth P. Diary keeping in asthma: comparison of written and electronic methods. BMJ 1993;306:487–9.
[6] Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. Public Opin Q 1996;60:275–304.
[7] Taenzer PA, Speca M, Atkinson MJ, et al. Computerized quality-of-life screening in an oncology clinic. Cancer Pract 1997;5:168–75.
[8] Bloom DE. Technology, experimentation, and the quality of survey data. Science 1998;280:847–8.
[9] Velikova G, Wright EP, Smith AB, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. J Clin Oncol 1999;17:998–1007.
[10] Stone AA, Shiffman S, Schwartz JE, et al. Patient non-compliance with paper diaries. BMJ 2002;324:1193–4.
[11] Bushnell DM, Reilly MC, Galani C, et al. Validation of electronic data capture of the Irritable Bowel Syndrome—Quality of Life Measure, the Work Productivity and Activity Impairment Questionnaire for Irritable Bowel Syndrome and the EuroQol. Value Health 2006;9:98–105.
[12] Tiplady B. Electronic patient diaries and questionnaires – ePRO now delivering on promise? Patient 2010;3:179–83.
[13] Yeomans A. The future of ePRO platforms. Appl Clin Trials January 28, 2014. Available from: www.appliedclinicaltrialsonline.com/appliedclinicaltrials/article/articleDetail.jsp?id=833920&pageID=1 [Accessed June 1, 2014].
[14] Ganser AL, Raymond SA, Pearson JD. Data quality and power in clinical trials: a comparison of ePRO and paper in a randomized trial. In: Byrom B, Tiplady B (eds.), ePRO: Electronic Solutions for Patient-Reported Data. Surrey, England: Gower, 2010.
[15] Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. (4th ed.). New York: Oxford University Press, 2008.
[16] Gnanasakthy A, DeMuro C, Boulton C. Integration of patient-reported outcomes in multiregional confirmatory clinical trials. Contemp Clin Trials 2013;35:62–9.
[17] Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value Health 2008;11:322–33.
[18] Rosenthal R. The "file drawer problem" and tolerance for null results. Psychol Bull 1979;86:638–41.
[19] Song F, Parekh-Bhurke S, Hooper L, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Med Res Methodol 2009;9:79.

[20] Ramachandran S, Lundy JJ, Coons SJ. Testing the measurement equivalence of paper and touch-screen versions of the EQ-5D visual analog scale (EQ VAS). Qual Life Res 2008;17:1117–20.
[21] Athale N, Sturley A, Skoczen S, et al. A web-compatible instrument for measuring self-reported disease activity in arthritis. J Rheumatol 2004;31:223–8.
[22] McCabe SE, Diez A, Boyd CJ, et al. Comparing web and mail responses in a mixed mode survey in college alcohol use research. Addict Behav 2006;31:1619–27.
[23] Bennett AV, Keenoy K, Basch E, Temple LK. Is between-mode equivalence comparable to test-retest reliability for patient-reported outcome (PRO) measures: a test case of Web versus IVRS versus paper for the MSKCC Bowel Function Instrument and LASA QOL. Value Health 2013;16:A33.
[24] Bennett AV, Dueck AC, Mitchell SA, et al. Mode equivalence and acceptability of Web-, Interactive Voice Response System-, and Paper-based administration of US National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Qual Life Res 2013;22:42.
[25] Lundy JJ, Coons SJ. Measurement equivalence of interactive voice response and paper versions of the EQ-5D in a cancer patient sample. Value Health 2011;14:867–71.
[26] Lundy JJ, Coons SJ, Aaronson NK. Testing the measurement equivalence of paper and interactive voice response system versions of the EORTC QLQ-C30. Qual Life Res 2014;23:229–37.
[27] Bent H, Ratzlaff CR, Goligher EC, et al. Computer administered bath ankylosing spondylitis and Quebec Scale outcome questionnaires for low back pain: agreement with traditional paper format. J Rheumatol 2005;32:669–72.
[28] Bellamy N, Campbell J, Stevens J, et al. Validation study of a computerized version of the Western Ontario and McMaster Universities VA3.0 Osteoarthritis Index. J Rheumatol 1997;24:2413–5.
[29] Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value Health 2009;12: (419–419).
[30] Raymond SA. Forward – the patient's viewpoint: impact of ePRO technologies on clinical research. In: Byrom B, Tiplady B (eds.), ePRO: Electronic Solutions for Patient-Reported Data. Surrey, England: Gower, 2010.
[31] Tiplady B. Diary design considerations: interface issues and patient acceptability. In: Byrom B, Tiplady B (eds.), ePRO: Electronic Solutions for Patient-Reported Data. Surrey, England: Gower, 2010.
[32] Taylor NP. Pfizer: 'Bring your own device' is coming to clinical trials. December 2, 2013. Available from: www.fiercebiotechit.com/story/pfizer-sees-smartphones-taking-center-stage-clinical-trials/2013-12-02. [Accessed June 1, 2014].
[33] Zbrozek A, Hebert J, Gogates G, et al. Validation of electronic systems to collect patient-reported outcome (PRO) data—recommendations for clinical trial teams: report of the ISPOR ePRO Systems Validation Good Research Practices Task Force. Value Health 2013;16:480–9.
[34] Willis GB. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks California: Sage, 2005.
[35] McEntegart D. Equivalence testing: validation and supporting evidence when using modified PRO instruments. In: Byrom B, Tiplady B (eds.), ePRO: Electronic Solutions for Patient-Reported Data. Surrey, England: Gower, 2010.
[36] Gallo P. Treatment-by-center interaction. In: D'Agostino R, Sullivan L, Massaro J (eds.), Wiley Encyclopedia of Clinical Trials. Hoboken: John Wiley & Sons, Inc., 2008.
[37] Wild D, Eremenco S, Mear I, et al. Multinational trials—recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient Reported Outcomes Translation & Linguistic Validation Good Research Practices Task Force report. Value Health 2009;12:430–40.