



# Methodological issues affecting the value of patient-reported outcomes data

Michael R Hufford<sup>†</sup> and Saul Schiffman

The validity and value of patient-reported outcomes data are heavily dependent on the methods used to collect the data. This review examines the impact of a variety of methodological issues on the value of patient-reported outcome data. In particular, when patients are asked to self-monitor their experiences, disease episodes and healthcare utilization over time, a variety of methodological issues must be addressed if the data are to be considered a reliable and valid reflection of their daily lives. Ecological momentary assessment, a set of methods for collecting real-time data from patients in their natural environments, holds considerable promise as a way to enhance the value of certain types of patient-reported outcome data.

*Expert Rev. Pharmacoeconomics Outcomes Res.* 2(2), (2002)

## CONTENTS

Methodological assumptions in PRO research

Moving from there-and-then, to here-and-now

Use of diaries to avoid recall

Ecological momentary assessment

Conclusions

Expert opinion

Five-year view

Key issues

Information resources

References

Affiliations

Pharmaceutical science has moved away from focusing exclusively on the impact of medications on physiological parameters to a focus that includes the patient's experiences, summarized as patient-reported outcomes (PROs). Although the ability of medications to treat disease is the primary focus of drug development, the ultimate goal of developing efficacious medicines is to improve patients' well-being. In short, patients increasingly demand not just to live longer and in better health by some objective medical standard, but also want to feel and function better. The increasing focus on PROs in medicine has had the positive effect of giving prominence to the views and experiences of patients [1].

The increasing interest in patients' experiences has been paralleled by a dramatic growth in the number and type of assessments used to collect PRO data. Hundreds of PRO measures now exist, spanning questionnaires assessing quality of life (QoL) in general, to disease- and symptom-specific questionnaires for very specific patient groups. The explosive growth in the number of assessments has elicited some criticism, as some have noted that race to create new PRO measures is sometimes driven by change for change's sake [2].

For all of the interest in the content of PRO assessments, relatively little attention has been paid to the methodological factors that can impact the reliability, validity and sensitivity of the data. If PRO data are to be highly valued, it needs to be a faithful reflection of the patient's experience, free from bias and noise that may obscure meaningful relationships to other variables of interest.

This review provides a brief background on methodological issues related to collection of PRO data. We focus particularly on one particular type of PRO data, diary data collected in patients' natural environments, as an example of the direct impact that methodology can have on the value of PRO data. A collection of methods that have evolved out of the behavioral sciences termed ecological momentary assessment (EMA) is presented as illustrating one recent development in the methodology of PRO research [3].

## Methodological assumptions in PRO research

This review will not attempt to catalog the spectrum of methodological factors that can affect PRO data collection, but rather draw on several illustrative examples. It is important to note that patients' self-reports are in general reliable. Moreover, for many symptoms and

<sup>†</sup> Author for correspondence  
Invivodata, Inc., 2100 Wharton  
St., Ste. 505, Pittsburgh, PA  
15203, USA  
Tel.: +1 412 390 3008  
Fax: +1 412 390 3020  
mhufford@invivodata.com

## KEYWORDS:

Ecological momentary assessment, methodology, recall, self-report

conditions (e.g., pain, fatigue), the required data can only be collected by asking the patient how they are feeling directly. The challenges when collecting PRO data are to emphasize the need to capture psychometrically sound data and to appreciate the impact that methodological factors can have on its validity and thus its value. Each step in the collection of PRO data can potentially impact the data. This review highlights five illustrative issues: the informed consent, the medium used to collect the data, setting effects, the cognitive load required to answer the typical PRO inquiry and reactivity to PRO assessment.

### ***Informed consent***

Even before the patient completes a PRO measure, what they have been told in the informed consent can impact their self-reports [4]. For example, the more patients understand the risks of disclosing sensitive information about their past behavior, the less likely they are to self-report that information [5]. Patients' awareness of what might happen as a consequence of their self-reports can also impact the data. In one study, patients were led to believe that the reporting of depressive symptoms could lead either to an intervention by the experimenter, including contact with the patient's significant other, or only minimal intervention. Patients in the intrusive intervention group reported significantly fewer symptoms of depression than patients led to believe that only minimal intervention was possible [6].

### ***Medium used to collect the data***

Whether the patient completes the PRO measure using paper-and-pencil or a computer can also affect the data. Some research suggests that patients disclose more sensitive data to a computer, believing the data are more anonymous than using paper-and-pencil [7]. Other studies confirm that patient will readily use office- and field-based electronic devices for data entry [8,9]. In terms of electronic diaries in particular, several studies have examined patient preference for electronic diaries and universally find that subjects readily accept and even prefer these devices to paper diaries [10,11]. Interestingly, patients' age, gender and comfort/familiarity with technology were not associated with their diary preference [12].

### ***Setting effects***

The setting in which the PRO data are collected can also impact the data. A good example of this effect is the well-documented phenomenon of 'white coat hypertension,' where blood pressures taken by 'white-coated' medical personnel in clinical settings are falsely elevated relative to actual real-world physiological parameters [13]. In other words, measures in artificial contexts (e.g., research sites) can fail to accurately capture the level and the range of responses observed in real-world settings.

### ***Cognitive load***

As others have noted, the validity of PRO data depends in large part on the clarity and precision of the question and on the patient's ability to answer it accurately [14]. The value of

patients' response to a PRO measure rests on their ability to understand the question and then perform the cognitive tasks required to generate an accurate, unbiased response. Although this cognitive task sounds straightforward, the typical PRO inquiry requires patients to step through a fairly complex set of cognitive operations to produce the requested data. We use the term cognitive load to reflect the amount and complexity of the cognitive processing required to answer a PRO inquiry.

For example, asking patients to report whether they are in pain right now presents a minimal cognitive load. The patient must introspect and make an appropriate response, but the cognitive effort required is fairly small: the data are essentially immediately and straightforwardly available to introspection. However, a common PRO inquiry, asking the patient to report how severe a symptom has been on average over the past 30 days, presents a much greater cognitive load. Responding to this inquiry accurately requires, at a minimum, the following steps:

- Recollection of the target symptom or event in question (which may imply enumeration of many occasions over a substantial time period)
- Anchoring their recall during the correct time period
- Aggregating experience of the symptom over time to arrive at a numeric average for the specific time period

When broken down into its component parts, the cognitive load required to answer this PRO inquiry is actually quite formidable. Indeed, research shows that patients do not even attempt such retrieval and summary, but shortcut the process by generating a heuristic – and highly bias-prone – impression of the experience in question. An empirical literature from cognitive psychology and a variety of methodological studies demonstrates that each of these steps can potentially adversely impact the accuracy and validity of the data.

### ***Recall***

Research on memory demonstrates that recall is not a straightforward retrieval of events. Memory relies on a variety of mental shortcuts, or heuristic strategies, to reconstruct past events [15]. This retrospective reconstruction is imperfect and subject to a range of biases resulting from the cognitive load required to answer the typical paper-and-pencil PRO inquiry. Each step in the process has the potential to introduce inaccuracy and bias into the data.

For example, in order for any content to be available for later recall, it must first be encoded into memory. Encoding is always incomplete and imperfect and is influenced by a variety of processes. For example, whether something is encoded at all is influenced by its perceived salience at the time, which may be a function of the person's subjective state or distraction by other stimuli [16,17]. Thus, autobiographical memory does not start with an objective, accurate record of events, but is both error- and bias-prone from the outset. When patients are asked to subsequently recall an encoded experience, they do not simply retrieve and 'replay' the experience in question. Rather, a variety of heuristics are used to retrieve the information. Among the heuristics known to impact the retrieval of information include:

- Availability: more salient, recent, or surprising, events can have an undue influence on recall [18,19]
- Saliency: The personal relevance of an experience can also affect encoding and retrieval, as salient, more intense experiences are more likely to be encoded initially and recalled later [20]
- Recency: More recent events are more accessible to memory and as a result exercise a disproportionate impact on recall [21,22]
- State biases: Like setting effects, state biases impact recall by virtue of making certain content more accessible. For example, people are more likely to retrieve negative information when they are in a negative mood, thus introducing substantial bias [23–25]
- Effort after meaning: People's natural and unconscious tendency is to reconstruct events so as to make them consistent with subsequent events, or to make them amenable to the construction of a coherent and comfortable story of the past [26–29]

This bias is particularly problematic because it tends to produce consistent and systematic results that often 'make sense' and conform to the research hypothesis. For example, one study found that the degree of fatigue reported by radiation therapy patients prior to treatment was actually influenced by whether their fatigue increased or decreased following radiation treatment. Thus, some patients overestimated prior fatigue and others underestimated it to make sense of their preradiation fatigue in light of their reaction to treatment [30].

#### Anchoring recall in time

Irrespective of the memory biases that may affect recall of PRO data, most researchers simply assume that patients can frame their experience in the requested timeframe (e.g., past week or month) with little effort. The anchoring of recall in time is one often-overlooked source of error. For example, Bailey and Martin analyzed patients' understanding of what time interval 'past month' referred to during completion of QoL questionnaires across 11 studies [31]. If patients were asked on the 15th of a month to recall the past month, only 64% of patients correctly interpreted 'past month' to refer to the period from the 15th of the current month to the 15th of the previous month. In other words, 36% used an incorrect interval as the basis of their recall (e.g., some patients used the past 2 weeks only, others ignored the previous 2 weeks and used the entire month preceding the current month and still other patients used a 6-week period). There is also strong evidence that memory for events is subject to 'telescoping' – a tendency to bring past events forward, closer to the current time – with the effect that recall of, say 'past month' events is likely to include events prior to that time [32]. In some studies, this bias could result in recall of experiences following treatment, including information about pretreatment status, seriously undermining evaluation of treatment effects. Additional research is needed to better understand how frequently recall anchoring errors occur in retrospective PRO measures.

#### Aggregating experience

The aggregation of experiences, once recalled, is also a source of potential bias. This type of cognitive processing is necessary to respond to questions about the occurrence or frequency of events ('How many headaches did you have this week?') or their average or typical characteristics ('How much gastric discomfort did you have this week?').

Findings from autobiographical memory research makes clear that patients do not process these requests by counting and numeric averaging, but rather they rely on a variety of heuristic strategies for estimating the answers. These heuristics potentially introduce substantial biases, similar to those that affect retrieval of information in the first place. For example, research shows that patients' recall of their 'average' pain is heavily influenced by the saliency (peak pain) and recency (pain intensity at the end of the episode) of the patient's pain experience during the recall interval [27,33,34]. Patients' most recent experiences can also affect their aggregation because estimates of the frequency of events are anchored by the effects of the most recent occurrences [15].

Furthermore, it is not evident that researchers are always explicit in their instructions to participants regarding aggregation. For example, when patients are asked to recall their pain level over the last week, are they being asked to recall their average pain level, or some other index of central tendency (e.g., their modal pain level)?

#### Reactivity to assessment

One challenge when collecting PRO data are its potential distortion because of reactivity. Reactivity is the degree to which the intensity, frequency and/or quality of a dependent variable changes as a function of being observed, monitored, or assessed [35]. Reactivity can be generated from at least two different sources – the content of the assessment itself and the methodology used to collect the data.

In terms of assessment content, reactivity presents researchers with a unique dilemma, namely that the content of the assessment can affect itself. For example, the Beck Depression Inventory (BDI) is a widely used self-report measure of depression severity [36,37]. In a series of studies, researchers examined whether the completion of the BDI affects self-reported mood, confounding the assessment of depression severity with that of induced mood. Results from the studies suggested that for nondepressed patients, completion of the BDI results in more positive affect, whereas depressed patients experienced more negative affect after completing the assessment [38].

#### Moving from there-and-then, to here-and-now

Many traditional paper-and-pencil PRO measures ask patients to retrospect about the occurrence, frequency and severity of a variety of different symptoms and events. In other words, patients are asked to report on the 'there-and-then' of their experience. One alternative to the reliance on recall data are to collect patients' experiences in the real world environments, in real-time. In other words, patients can also be asked to report on their experiences in the 'here-and-now.'

One common type of there-and-then data involves the evaluation of patients' QoL. One ramification of the reliance on intermittent, retrospective reports of QoL is that they are vulnerable to response shifts, or 'treadmill' effects [39]. Patients' reports of well-being change depending on what they are comparing it to (i.e., their well-being baseline). As well-being changes over time, patients' reports of well-being can become dependent on a fluctuating baseline, resulting in a 'satisfaction treadmill' effect. As a patient experiences improvement of a condition, the new, improved well-being becomes the new baseline and objective improvements in their well-being may not result in increased well-being reports because they have become accustomed to the new, higher level of well-being. The net effects of the satisfaction treadmill is that global summary reports of QoL may fail to yield significant differences over time and treatment despite actual improvements in patients' QoL.

The subcomponents of QoL, from subjective well-being to objective functioning, derive from an interest in understanding a patient's daily and momentary experiences. In other words, a summary judgment that symptoms are less severe or that daily functioning is improved rests on the accumulation of moment-to-moment experiences. Rather than relying on a 'top-down,' broad summary of momentary experience, a 'bottom-up' approach to QoL assessment directly measures momentary experience and changes as a function of treatment or illness progression over time. That is, if we are interested in sensitively measuring changes in QoL over time, we are fundamentally interested in sampling patients' daily and momentary experiences [40]. Rather than being at odds with traditional retrospective measures of QoL, this 'bottom-up,' here-and-now approach provides a complimentary perspective on the extent to which patients' evaluations of various moments in their lives reflect QoL changes over time. Besides providing a more detailed and discriminating view of PRO data, this approach avoids the inaccuracies and biases that affect recall data.

#### Use of diaries to avoid recall

Historically, researchers have used paper diaries as a common method for collecting 'here-and-now' data from patients in the field. Initially, diary entries were meant to cue patients' subsequent recall of health events and were not seen as a primary source of data [41]. Over time, paper diaries became a way to capture recall and summary data over a relatively brief interval, such as a day, in an attempt to reduce recall biases.

Many studies use paper diaries to collect PRO data once daily from patients [42,43]. Given that retrospective questionnaires often ask for recall over intervals of weeks or months, daily summaries can potentially provide substantial benefit in bringing data collection closer to real-time and minimizing recall. Indeed, for some types of events and recall tasks, daily summary recall may well be adequate. For example, it is likely that patients can recall singular and prominent events, such whether or not they had a migraine headache at any time during the day. On the other hand, for some data even brief periods of recall can dramatically affect the results [33]. For example, Stone

*et al.* [44] have shown that patients' 24–48 h recall of how they coped with difficulties is subject to substantial error (patients forget a substantial proportion of their coping responses) and bias (cognitive strategies are especially likely to be omitted).

Two other methodological problems prevent paper diaries from being an effective way to capture reliable and valid PRO data. First, data quality is notoriously problematic, with many entries being illegible or containing out-of-range data. Perhaps more importantly, there is no way to verify whether or not patients are actually completing their diary cards as required by the protocol. A small body of empirical literature suggests that paper diaries are often not completed as required [45], but rather 'hoarded' and completed before a site visit, so-called 'parking lot compliance.' Patient noncompliance with paper diary protocols means that the data are now subject to the very recall biases that were the rationale for a diary study in the first place [46].

#### Ecological momentary assessment

A number of methods have been developed to acquire patient experience data with less distortion than is found using recall methodologies and fewer data quality and noncompliance problems than is found using paper diaries. The term EMA refers to a conceptual strategy of collecting real-time momentary data from patients in real-world settings [3]. We first describe the key characteristics of EMA, including how different types of sampling can be used to tailor the research protocol to match the study objectives. Next, the technology that can be used to implement EMA is presented. Finally, we discuss some of the challenges in implementing EMA designs.

#### Characteristics of EMA

Three characteristics that define EMA: ecological validity, a momentary focus and repeated assessments [3]. By putting these facets together using a combination of sampling strategies, researchers can have a unique perspective on patients' real-time experiences.

#### Ecological validity

In terms of ecological validity, patients are studied in the environments they typically inhabit and under conditions that typify their daily life, to maximize real-world generalization. Indeed, most clinical research aims to understand real world patient experiences as they occur in the context of everyday life. The ecological focus of EMA is in contrast to laboratory- or clinic-based studies, where patients are assessed in settings considerably different from those of their day-to-day lives.

One common real world influence on PROs is symptom severity. When patients are asked to look retrospectively at a particular symptom, such as pain severity, they must place that recall in the context of everyday life. If our interest is in understanding the impact of a treatment on pain severity, it is more ecologically valid to collect pain data at the point of experience (e.g., after having vacuumed the house), rather than in a pain clinic. The corresponding loss of ecological validity when all

data are collected in-clinic deprives researchers of the ability to sensitively tests for symptom severity improvements across treatments. In sum, EMA maximizes ecological validity by allowing respondents to report on their experiences in their real world environment.

#### Momentary focus

EMA also focuses on collecting data at or near real-time. By focusing on near-immediate experiences, retrospective distortion of self-report data by recall processes is minimized, which helps to maximize the validity of the reports and ultimately enhance the value of the data. Real-time measures place less of a cognitive load on patients. By reporting their current experiences and behaviors, patients' real-time reports are more likely to be a faithful reflection of their actual experiences.

#### Repeated assessments

EMA's approach to understanding PRO data emphasizes capturing both representative and unique moments in a patient's life. In a typical EMA study, patients' momentary experience is repeatedly assessed as they go about their daily activities. For example, patients might be 'beeped' at random intervals to assess their well-being. Implementing intensive momentary assessments requires careful attention to psychometrics and patient compliance. Assessments must be geared toward assessing momentary PROs and be easy to use and administer. Carefully constructed EMA measures have demonstrated good reliability and validity [47].

Just as thoughtful consideration of which patients to include in a sample is critical to drawing valid conclusions, so too is the careful, unbiased sampling of moments critical to drawing valid conclusions about PROs. Designing a study that appropriately samples moments in a patient's life ensures maximum sensitivity to treatment effects by capturing data on the frequency, intensity and duration of momentary, here-and-now, experiences.

Collecting PRO data at multiple time-points also enables researchers to examine changes in patients' experiences over time. Rarely is symptom severity static: rather it ebbs and flows as a function of circadian rhythms (e.g., diurnal variation in pain severity) and time since treatment (e.g., fatigue secondary to chemotherapy). Global retrospective measures of symptom severity in a doctor's office can fail to take into account the temporal dynamics known to affect patients' symptom severity. Repeatedly assessing patients over time ensures a reasonable characterization of the phenomenon under study and enables researchers to examine fluctuations in the phenomenon over time. Various methods of analyzing data have been employed to characterize EMA data [48].

#### Sampling

EMA typically uses a combination of sampling strategies to capture real-time data from patients over time. Many EMA sampling strategies involve a combination of several basic types of assessment contingencies: event-, interval- and signal-contingent [49].

#### Event-contingent

Event-contingent sampling relies on the patient to self-monitor when a target experience occurs and then to complete an assessment. Such data may be used to simply count events or to collect data about events, such as episodes of urinary incontinence or migraine headaches. This type of contingency ensures that events are not missed, but requires vigilance to constantly monitor for the specific experience and compliance with the real-time reporting. As outlined above, in some circumstances self-monitoring using event-contingent paradigms can induce reactive effects.

#### Interval-contingent

Interval-contingent recording requires patients to complete an assessment at a preselected time (e.g., end of day) or times of the day (e.g., 10am, 2pm, 8pm). This type of contingency can be used to ensure that specific times during the day are sampled and can also be used to prompt or remind the patient to engage in certain types of behaviors, like medication taking. Limitations of this type of contingency include: that it may miss episodic events, that the prescribed schedule may not be representative of patient's lives (especially end of day sampling) and that the timing of the assessment may be anticipated, which could lead to response biases. Broadly speaking, these drawbacks argue against the implementation of exclusively interval-contingent sampling strategies.

#### Signal-contingent

Signal-contingent sampling requires the use of some type of prompting device (e.g., handheld computer) to cue the patient to complete an assessment. The exact schedule of prompting can be driven by the needs of the protocol. For example, signal-contingent reports are often issued at random intervals to collect representative data from patients. More dynamic sampling can be employed to over-sample critical periods; for example, to detect a medication effect hypothesized to be greatest at a particular time of day [50]. The limitations of signal-contingent recording include the reliance on patient compliance with the prompting schedule and like interval-contingent recording, they may miss important episodes. However, signal contingent approaches can be combined with event-contingent recording to capture both events and average states [9].

#### Implementing EMA

The benefits of EMA, including the reduced cognitive load of the questions posed to patients, are balanced by the technological requirements to execute these studies. In general, the cognitive load of PRO assessments tends to be inversely related to the methodological burden placed on the researcher. As the assessments require less and less cognitive effort to complete, the methodology and technology used to collect the data must become increasingly sophisticated.

Most EMA protocols use some type of technological device to prompt patients to complete their assessments, at a minimum. For example, 'smart' wristwatches have been used to

prompt patients. However, both pagers and 'smart' wrist-watches often have to limit their data collection to specific windows during the day, often between 8am and 8pm, to avoid waking patients with varying sleep habits. This is not trivial, as studies have found that up to 26% of episodes fall outside an 8am to 8pm time window and assessments outside of this window differ significantly from assessments within it [51]. Another limitation of these signaling devices is that they confirm only that a prompt was issued, not that the report was completed in a timely way. For example, a study using programmable wristwatches by Litt *et al.* suggested that patients often completed assessments long after they were beeped [45]. This undermines the purpose of executing an EMA study, as the data are now subject to the recall biases that motivated an EMA study in the first place.

For more than a decade, researchers have used commercially available handheld computers as a way to implement EMA designs and collect e-PRO data [46]. As electronic prompting devices, their programmability allows flexibility to implement a variety of research and sampling designs, including those where the prompting and sampling protocol changes dynamically (e.g., prompting more frequently during an episodic increase in symptoms). Their ability to turn prompting off dynamically for sleep or other 'time-out's is also valuable and appreciated by patients. Sophisticated electronic diaries can also prompt and track compliance with other aspects of the protocol, such as medication taking. Moreover, they can also be designed to promote compliance, for example, by providing patients with real-time feedback about their compliance to the protocol or medication regimens [46].

As assessment platforms, sophisticated electronic diaries also provide researchers with several advantages over paper-based diaries. All patient entries and interactions with PEDs can be time-and-date stamped, allowing for a verifiable electronic record of their compliance with the protocol. This is not a trivial matter, as many anecdotal and quantitative reports suggest that patients hoard their diaries and complete them in batches when they are due at the research site [52–59]. The user interface can be developed so that it is easy to use, further reducing the cognitive load of the assessments. Modern interfaces allow point-and-touch interfaces that can be used by a wide variety of patients [60]. Since the assessments are computerized, it becomes possible to build in edit checks to ensure in-range data and completion of obligatory fields.

One example of a recent EMA study was a clinical trial comparing the efficacy of two different nicotine replacement patches for smoking cessation [50]. The key momentary variables of interest included the experience of craving and withdrawal symptoms, especially during the morning hours. Smokers were taught to use a PED to record details regarding their craving, withdrawal and smoking behavior in real-time over a 3-week period. The PEDs allowed for both patient-initiated entries (e.g., significant smoking temptation episodes), as well as 8–10 signal-contingent assessments on a stratified random

schedule. Patients were highly compliant to the PED protocol, completing 93% of their signal-contingent assessments within 2 min of the audible prompt. By dynamically over-sampling the morning period, this study was able to distinguish between two types of nicotine replacement patches in terms of craving intensity and frequency, temptations to smoke and nicotine withdrawal symptoms (e.g., feeling anxious, irritable, or restless). Thus, technological advances in EMA diary platforms can be used to serve scientific and research design objectives.

### Challenges of implementing EMA

The challenges to implementing EMA designs include issues regarding the technology, as well as methodological challenges when dealing with real-time data.

The advantages to using handheld computers to implement EMA designs must be balanced against their potential drawbacks. There is a technological and financial burden attendant on conducting EMA studies using the highest level of technological innovation. Moreover, it is clear that technology alone is insufficient to obtain high rates of patient compliance with the protocol. For example, several studies that have not adequately addressed the user interface and other programming issues saw high rates of technical failure and patient noncompliance [45,61].

Another potential challenge that faces all types of PRO data collection is the potential distortion due to reactive effects. Three studies have examined whether EMA engenders significant reactivity among patients. All three studies failed to find evidence that EMA produces significant reactivity [62–64]. Regardless of this, it is important that research examine the conditions under which reactive effects may emerge and how they can be controlled.

EMA studies that use prompting (especially at random intervals) also pose reporting challenges for certain types of patients, such as those with occupations where it is difficult to respond to a signaled assessment (e.g., airline pilot or surgeon). Fortunately, PED studies that have monitored compliance have generally shown very high compliance rates (many over 90% [50,65]) with diverse samples, suggesting that inability to respond may not be a serious threat.

More broadly, EMA data can pose data management and analysis challenges. Firstly, a relatively simple EMA design that prompts 100 patients 5 times per day for a brief 10-item assessment, along with an average of 3 additional patient-initiated 'event' reports per day, over a 4-week period will produce approximately 224,000 data points. Some researchers may choose to create aggregates of patient experience by collapsing across multiple assessments each day because these aggregates will now be free of the retrospective inaccuracies and biases associated with recall data. Luckily, these aggregates lend themselves to traditional parametric analyses. More sophisticated analytic approaches, like generalized estimating equations and hierarchical regression analyses, have also appeared in the literature and are now available in commercial statistical packages [48,66].

## Conclusions

A combination of forces is driving an increased interest in PRO data. An aging, educated population, increasing competition among drugs with similar efficacy profiles and direct-to-consumer advertising are a few of the factors behind the burgeoning interest in assessing PROs. Researchers are also driving this interest in PRO data through the development of an ever-increasing number of PRO assessments. As more pressure is placed on researchers to justify the value of their PRO data, sources of error and bias are increasingly being examined to enhance the predictive value of these data. A number of methodological factors can impact the value of PRO data. Perhaps most surprisingly, it is often the overlooked aspects of how a PRO measure is implemented that can be the source of error and bias. From the informed consent, to the setting of the measure administration, through to the impact of the assessments content on the very phenomena that is being assessed, methodological factors play an important role in determining the reliability, validity and ultimately the value of PRO data.

The inaccuracies and biases associated with recall were identified as one of the prime contributors to error and bias. Asking patients to recall details regarding their experience places a considerable cognitive load on them to faithfully recall the target symptom or experience, anchor their recall appropriately and then aggregate their experience to produce the requested data. As the cognitive load of a measure increases, the potential for well-documented recall biases to affect the data increases, threatening the validity and thus the value, of the resulting data. One alternative to the reliance on recall for PRO data are to construct a 'bottom-up' approach, where patients' experiences are sampled in real-time. A collection of methods that has evolved from the behavioral sciences, termed EMA, is one way to collect real-time, real-world data from patients. The application of EMA to PRO research highlights a basic principle – the cognitive load required of patients to answer a PRO inquiry tend to vary inversely with the amount of the methodological burden placed on the researcher. Thus, the technology requirements of sophisticated EMA solutions is much greater than the simplicity of

asking patients to recall their experience over time during a research site visit. As was outlined above, the potential value of EMA data can outweigh the methodological requirements depending on the nature of data that is being collected.

## Expert opinion

The myriad of PRO assessments is matched only by the diversity of experiences and outcomes that they are trying to measure. When researchers are interested in PROs that are rooted in day-to-day or moment-to-moment experiences, EMA provides an unparalleled way to collect valid data, free from the biases affecting recall data. Currently, 25% of clinical trials collect diary data from patients, reflecting researchers' pervasive interests in the impact of medications on patients' real-world experiences [22]. Whenever studies focus on symptoms, behaviors and events that occur in real life, EMA provides researchers with a way to collect sensitive data from patients, while reducing the cognitive load of the assessments.

## Five-year view

The race to develop new PRO measures will continue, with a gradual shift in attention away from a sole focus on the content of measures, to one that includes methodological factors that can affect data. Over the next 5 years, researchers will become increasingly cognizant of the impact of methodological factors on the reliability, validity and ultimately the value of PRO data. The pernicious biases associated with reliance on recall will help focus research attention on methodologies that help avoid these biases and reduce the cognitive load of PRO assessments on patients. Importantly, researchers will also become more comfortable with the application of technology in the service of enhancing the value of PRO data [67]. The use of EMA in academic research and clinical trials will increase over time, becoming one common way to collect PRO data when the research question involves an understanding of patients' real world experiences [68,69]. More broadly, EMA will be used as the gold standard against which summary, retrospective measures of PROs are validated.

## Key issues

- A variety of methodological factors can affect the value of patient-reported outcome (PRO) data.
- Key methodological factors that are often overlooked as a source of error and bias in PRO data include: the informed consent process, the medium used to collect the data, setting effects, the cognitive load required by the patient to generate the data and reactivity to assessment.
- Real-time measures of PROs place less cognitive load on patients than do summary, recall measures.
- Many PROs have their foundations in moment-to-moment experiences, placing a premium on the use of diaries to collect the data.
- In general, there is an inverse relationship between the cognitive load required by patients to produce PRO data and the methodological burden placed on researchers to collect the data.
- Ecological momentary assessment evolved out of the behavioral sciences as way to capture real-time, real world data from subjects, free from the biases associated with recall [3].

## Information resources

For a current summary of factors affecting the value of self-report data, read:

- Stone AA *et al.* *The Science of Self-Report: Implications for Research and Practice*. Lawrence Erlbaum Associates, Mahwah, NJ, USA (2000).

For a review of the key aspects of EMA, read:

- Stone AA, Shiffman S. Ecological Momentary Assessment: Measuring real world processes in behavioral medicine. *Ann. Behavioral Med.* 16, 199–202 (1994).

For a review of the development of EMA and patient compli-

ance with diaries in clinical trials, read:

- Hufford MR, Shiffman S, Paty J, Stone AA. Ecological momentary assessment: Real world, real-time measurement of patient experience. In: Progress in Ambulatory Assessment, Fahrenberg J, Myrtek M (Eds.) Hogrefe & Huber Publishers, Seattle, WA, USA, 69–92 (2001).

For a summary of the applications of electronic diaries to the collection of EMA data in clinical trials, read:

- Shiffman S, Hufford MR. Subject experience diaries in clinical research, Part 2: Ecological Momentary Assessment. *Applied Clin. Trials* 10, 42–48 (2001).

## References

Papers of special note have been highlighted as:

- of interest
- of considerable interest

- 1 Lepage A, Hunt S. The Problem of Quality of Life in Medicine. *J. Am. Med. Assoc.* 278, 47–50 (1998).
- 2 Kind P. Measuring up to the task. *Value in Health* 4, 344–345 (2001).
- 3 ••Stone AA, Shiffman S. Ecological Momentary Assessment: Measuring real world processes in behavioral medicine. *Ann. Behavioral Med.* 16, 199–202 (1994).
- 4 Bersoff DM, Bersoff DN. Ethical issues in the collection of self-report data. *The Science of Self-Report: Implications for Research and Practice*. Stone AA, Turkkan JS, Bachrach CA, Jobe JE, Kurtzman HS, Cain VS. (Eds.) Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA, 9–24 (2000).
- 5 Singer E, Frankel MR. Informed consent procedures in telephone interviews. *Am. Sociological Rev.* 47, 416–427 (1982).
- 6 Stanton AL, Burkner EJ, Kershaw D. Effects of researcher follow-up on distressed subjects; Tradeoff between validity and ethical responsibility? *Ethics & Behavior* 1, 105–112 (1991).
- 7 Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use and violence: increased reporting with computer survey technology. *Science* 280, 847–848 (1998).
- 8 Göbel H, Heinze A, Kuhn K, Heuss D, Linder V. Effects of operationalized computer diagnosis on the therapeutic results of sumatriptan in general practice. *Cephalgia* 18, 481–486 (1998).
- 9 Shiffman S, Paty JA, Gnys M, Kassel JA, Hickcox M. First lapses to smoking: Within-subjects analysis of real-time reports. *J. Consult. Clin. Psychol.* 64, 366–379 (1996).
- 10 Drummond HE, Ghosh S, Ferguson A, Brackenridge D, Tiplady B. Electronic quality of life questionnaires: A comparison of pen-based electronic questionnaires with convention paper in a gastrointestinal study. *Qual. Life Res.* 4, 21–26 (1995).
- 11 •Rabin JM, McNett J, Badlani GH. ‘Compu-voiding II’: The computerized voiding diary. *J. Med. Syst.* 20, 19–34 (1996).
- 12 •Tiplady B, Crompton GK, Dewar MH *et al.* The use of electronic diaries in respiratory studies. *Drug Info. J.* 31, 759–764 (1997).
- 13 Pickering TG, Coats A, Mallion JM, Mancia G, Verdecchia P. Blood pressure monitoring. Task force V: White-coat hypertension. *Blood Pres. Monit.* 4, 333–341 (1999).
- 14 Baldwin W. Information no one else knows: The value of self-report. *The Science of Self-Report: Implications for Research and Practice*. Stone AA, Turkkan JS, Bachrach CA, Jobe JE, Kurtzman HS, Cain VS. (Eds.), Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA, 3–7 (2000).
- 15 ••Bradburn NM, Rips LJ, Shevell SK. Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science* 236, 157–161 (1987).
- 16 Erickson JR, Jemison CR. Relations among measures of autobiographical memory. *Bull. Psychonomic Soc.* 29, 233–236 (1991).
- 17 Linton SJ. Memory for chronic pain intensity: Correlates of accuracy. *Percept. Mot. Skills* 72, 1091–1095 (1991).
- 18 Eisenhower D, Mathiowetz NA, Morganstein D. Recall error: Sources and bias reduction techniques. In: *Measurement errors in surveys*. Beimer P, Groves R, Lyberg L, Mathiowetz N, Sudman S (Eds.), Wiley, New York, USA (1991).
- 19 Joffe RT, MacDonald C, Kutcher SP. Life events and mania: a case-controlled study. *Psychiatry Res.* 30, 213–216 (1989).
- 20 •Menon G, Yorkton EA. The Use of Memory and Contextual Cues in the Formation of Behavioral Frequency Judgments. In: *The Science of Self-Report: Implications for Research and Practice*. Stone AA, Turkkan JS, Bachrach CA, Jobe JE, Kurtzman HS, Cain VS. (Eds.), Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA, 63–80 (2000).
- 21 Schwarz N, Sudman S. *Autobiographical memory and the validity of retrospective reports*. Springer-Verlag, New York, USA (1994).
- 22 •Shiffman S, Hufford M, Hickox M *et al.* Remember that? A comparison of real-time versus retrospective recall of smoking lapses. *J. Consult. Clin. Psychol.* 65, 292–300 (1997).
- 23 Clark DM, Teasdale JD. Diurnal variation in clinical depression and accessibility of memories of positive and negative experiences. *J. Abnormal Psychol.* 91, 87–95 (1982).
- 24 Clark DM, Teasdale JD. Constraints on effects of mood on memory. *J. Personality Soc. Psychol.* 48, 1595–1608 (1985).
- 25 Teasdale JD, Fogarty SJ. Differential effects of induced mood on retrieval of pleasant and unpleasant events from episodic memory. *J. Abnormal Psychol.* 88, 248–257 (1979).
- 26 Brown GW, Harris T. *Social Origins of Depression*. Tavistock, London, UK (1978).
- 27 Eich E, Reeves JL, Jaeger B, Graff-Radford SB. Memory for Pain: Relation between Past and Present Pain Intensity. *Pain* 23, 375–379 (1985).
- 28 Means B, Swan GE, Jobe JB, Esposito JL. The Effects of Estimation Strategies on the Accuracy of Respondents’ Reports of

- Cigarette Smoking. *Autobiographical Memory and the Validity of Retrospective Reports*. Schwartz N, Sudman S. (Eds), Springer-Verlag, New York, USA, 107-119 (1994).
- 29 Schwartz ED, Kowalski JM, McNally RJ. Malignant memories: Post-traumatic changes in memory in adults after a school shooting. *J. Traumatic Stress* 6, 545-553 (1993).
- 30 Sprangers MAG, Van Dam FS, Broersen J *et al.* Revealing response shift in longitudinal research on fatigue: The use of the thentest approach. *Acta Oncology* 38, 709-718 (1999).
- 31 Bailey AS, Martin ML. In quality of life questionnaires, what do patients understand 'past month' to mean? Poster presented at the International Society for Quality of Life Research conference, Vancouver, Canada (October, 2000).
- 32 ••Bradburn NM. Temporal Representation and Event Dating. *The Science of Self Report: Implications for Research and Practice*. Stone AA, Turkkan JS, Bachrach CA, Jobe JE, Kurtzman HS, Cain VS. (Eds.), Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA, 49-63 (2000).
- 33 ••Redelmeier D, Kahneman D. Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66, 3-8 (1996).
- 34 Stone AA, Broderick JB, Kaell AT, Delespaul P, Porter L. Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritis? *J. Pain* 1, 203-218 (2000).
- 35 Nelson RO. Assessment and therapeutic functions of self-monitoring. In: *Progress in Behavioral Modification*. Hersen M, Eisler RM, Miller P (Eds), Academic Press, New York, USA, 3-41 (1977).
- 36 Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch. Gen. Psychiatry* 4, 561-571 (1964).
- 37 Beck AT, Steer RA, Brown GK *Beck Depression Inventory-II*. San Antonio, TX, The Psychological Corporation (1996).
- 38 Mark MM, Sinclair RC, Wellens TR. The effect of completing the Beck Depression Inventory on self-reported mood state: Contrast and assimilation. *Personal. Soc. Psychol. Bull.* 17, 457-465, (1991).
- 39 •Kahneman D. Objective Happiness. In: *Well-Being: The Foundations of Hedonic Psychology*. Kahneman D, Deiner E, Schwarz N (Eds.) Russell Sage Foundation, New York, USA, 3-25 (1999).
- 40 Barge-Schaapveld DQ, Nicolson NA, Berkhof J, deVries MW. Quality of life in depression: Daily life determinants and variability. *Psychiatry Res.* 88, 173-189 (1999).
- 41 Verbrugge LM. Health diaries. *Med. Care* 18, 73-95 (1980).
- 42 Lewis B, Lewis D, Cumming G. The Comparative Analgesic Efficacy of Transcutaneous Electrical Nerve Stimulation and a Non-steroidal Antiinflammatory Drug for Painful Osteoarthritis. *Br. J. Rheumatology* 33, 455-460 (1994).
- 43 Neugebauer R. Reliability of seizure diaries in adult epileptic patients. *Neuroepidemiology* 8, 228-233 (1989).
- 44 Stone AA, Schwartz JE, Neale JM *et al.* A comparison of coping assessed by ecological momentary assessment and retrospective recall. *J. Personal. Soc. Psychol.* 74, 1670-1680 (1998).
- 45 Litt MD, Cooney NL, Morse P. Ecological momentary assessment (EMA) with treated alcoholics: Methodological problems and potential solutions. *Health Psychol.* 17, 48-52 (1998).
- 46 •Hufford MR, Shiffman S, Paty J, Stone AA. Ecological momentary assessment: Real world, real-time measurement of patient experience. In: *Progress in Ambulatory Assessment*. Fahrenberg J, Myrtek M (Eds.), Hogrefe & Huber Publishers, Seattle, WA, USA, 69-92 (2001).
- 47 Stone AA, Shiffman S, DeVries MW. Ecological Momentary Assessment. In: *Well-Being: The foundations of hedonic psychology*. Kahneman D, Diener E, Schwarz N (Eds.), Russell Sage Foundation, New York, USA, 26-39 (1999).
- 48 Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. *Health Psychol.* 17, 6-16 (1998).
- 49 Wheeler L, Reis HT. Self-recording of everyday life events: Origins, types and uses. *J. Personality* 59, 339-354 (1991).
- 50 ••Shiffman S, Elash CE, Paton SM *et al.* Comparative efficacy of 24-hour and 16-hour transdermal nicotine patches for relief of morning craving. *Addiction* 95, 1185-1195 (2000).
- 51 Shiffman S. Real-time self-report of momentary states in the natural environment: Computerized ecological momentary assessment. In: *The Science of Self Report: Implications for Research and Practice*. Stone AA, Turkkan JS, Bachrach CA, Jobe JE, Kurtzman HS, Cain VS. (Eds.), Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, USA, 277-296 (2000).
- 52 Chmelik F, Doughty A. Objective measurements of compliance in asthma treatment. *Ann. Allergy* 73, 527-532 (1994).
- 53 Jonasson G, Carlsen K, Sodal A, Jonasson C, Mowinckel P. Patient compliance in a clinical trial with inhaled budesonide in children with mild asthma. *Eur. Respir. J.* 14, 150-154 (1999).
- 54 Mazze RS, Shamoan H, Pasmantier R *et al.* Reliability of blood glucose monitoring by patients with diabetes mellitus. *Am. J. Med.* 77, 211-217 (1984).
- 55 Milgrom H, Bender B, Ackerson L, Bowry P, Smith B, Rand C. Noncompliance and treatment failure in children with asthma. *J. Allergy Clin. Immunol.* 98, 1051-1057 (1996).
- 56 Simmons MS, Nides MA, Rand CS, Wise RA, Tashkin DP. Unpredictability of deception in compliance with physician prescribed bronchodilator inhaler use in a clinical trial. *Chest* 118, 290-295 (2000).
- 57 Spector S, Kinsman R, Mawhinney H *et al.* Compliance of patients with asthma with an experimental aerosolized medication: Implications for controlled clinical trials. *J. Allergy Clin. Immunol.* 77, 65-70 (1986).
- 58 Straka R, Fish J, Benson S, Suh J. Patient self-reporting of compliance does not correspond with electronic monitoring: An evaluation using isosorbide dinitrate as a model drug. *Pharmacotherapy* 17, 126-132 (1997).
- 59 Verschelden P, Cartier A, L'Archeveque A, Trudeau C, Malo JL. Compliance with and accuracy of daily assessment of peak expiratory flows (PEF) in asthmatic subjects over a three month period. *Eur. Respir. J.* 9, 880-885 (1996).
- 60 Shiffman S, Hufford MR, Paty J. Subject

- experience diaries in clinical research, Part 1: The patient experience movement. *Appl. Clin. Trials* 10, 46–56 (2001).
- 61 Johannes CB, Crawford SL, Woods J *et al.* An electronic menstrual cycle calendar: comparison of data quality with a paper version. *Menopause* 7, 200–208 (2000).
- 62 Cruise CE, Broderick J, Porter L, Kaell AT, Stone AA. Reactive effects of diary self-assessment in chronic pain patients. *Pain* 67, 253–258 (1996).
- 63 Hufford MR, Shields AL, Shiffman S, Paty J, Balabanis M. Reactivity to Ecological Momentary Assessment: An Example Using Undergraduate Problem Drinkers. *Psychol. Addict. Behav.* (In Press).
- 64 Peters ML, Sorbi MJ, Kruse DA, Kerssens JJ, Verhaak PF, Bensing JM. Electronic diary assessment of pain, disability and psychological adaptation in patients differing in duration of pain. *Pain* 84, 181–192 (2000).
- 65 Kamarck TW, Shiffman SM, Smithline L *et al.* Effects of task strain, social conflict and emotional activation on ambulatory cardiovascular activity: Daily life consequences of recurring stress in a multiethnic adult sample. *Health Psychol.* 17, 17–29 (1998).
- 66 Tennen H, Affleck G. Daily Processes in coping with chronic pain: Methods and analytic strategies. In: *Handbook of Coping: Theory, Research, & Applications* Zeidner M, Endler NS (Eds.), John Wiley & Sons, Inc., New York, USA, 151–177 (1996).
- 67 Hufford MR, Shields AL. Electronic subject diaries: An examination of applications and what works in the field. *Appl. Clin. Trials* (under review).
- 68 •Shiffman S, Stone AA. Introduction to the special section: Ecological momentary assessment in health psychology. *Health Psychol.* 17, 3–5 (1998).
- 69 Shiffman S, Hufford MR. Subject experience diaries in clinical research, Part 2: Ecological Momentary Assessment. *Appl. Clin. Trials* 10, 42–48 (2001).

**Affiliations**

- Michael R Hufford, Director, Scientific Affairs, Invivodata, Inc., 2100 Wharton St., Ste. 505, Pittsburgh, PA 15203, USA, Tel.: +1 412 390 3008, Fax: +1 412 390 3020, mbufford@invivodata.com
- Saul Shiffman, Chief Science Officer, invivodata, inc., Professor of Psychology (Health and Clinical), Psychiatry, & Pharmaceutical Sciences, University of Pittsburgh, 2100 Wharton St., Ste. 505, Pittsburgh, PA 15203, USA, Tel.: +1 412 390 3000, Fax: +1 412 390 3020, saul@invivodata.com