Draft Guidance for Industry
Patient-Reported Outcome Measures:
Use in Medical Product Development to Support Labeling Claims
Docket No. 2006D-0044

Submitted by invivodata, inc.

Saul Shiffman, Ph.D., Chief Science Officer
Jean Paty, Ph.D., Executive Vice President, Scientific Affairs

We commend FDA for their carefully-considered Draft Guidance addressing the important issue of Patient-Reported Outcomes (PROs) in clinical development. The issuance of the Draft Guidance reflects the recognition that PRO data frequently play critical roles in the evaluations of drugs, biologics, and devices (Willke, Burke & Erickson, 2004), and not only serve as the basis for labeling claims, but also as primary end-points for approval. The Draft Guidance and its focus on standards for development and validation of PRO assessments and procedures recognizes the well-developed science of self-report that underlies PRO assessment. The Draft Guidance will provide sponsors and other scientists with guidelines that will improve the application of PROs in clinical development.

Below, we comment on the Draft Guidance, noting both issues that we believe are well-supported by the scientific literature, as well as areas where the Guidance might be clarifies or strengthened to reflect the scientific evidence and best practices.

Recommendation to avoid relying on recall 2

Documenting timely, in-protocol completion of diary entries 4

Assessing actual behavior rather than patient beliefs, where possible 5

Implied standards and burden for migrating assessments from print to text-based electronic format should be based on scientific evidence. 6

Specific concerns when using electronic PRO instruments 9

Concerns about missing data 10

Respondent burden and automated skip patterns 11

Pooling of data from different assessment modalities. 12

Assessment frequency and timing 13

References 14

Recommendation to avoid relying on recall

In Lines 329-330, the Draft Guidance argues against over-reliance on recall, noting that "*it is important to consider patients' ability to accurately recall the information requested as proposed*." The Draft Guidance further notes that "PRO instruments that require patients to rely on memory, especially if they must recall over a period of time, or to average their response over a period of time may threaten the accuracy of the PRO data. It is usually better to construct items that ask patients to describe their current state than to ask them to compare their current state with an earlier period or to attempt to average their experiences over a period of time" (Draft Guidance, Lines 339-343). We agree with this position, and believe that it is consistent with a substantial body of evidence showing that self-reports can be systematically biased by the use of long recall periods.

Although random noise is a problem for the reliability of PROs and consequentially the statistical power of study designs, systematic biases deserve particular attention, because they can distort findings, for example creating apparent group differences where none exist or obscuring true group differences. The scientific literature from diverse areas such as cognitive science, autobiographical memory, and survey methodology supports the Agency's concern about relying on recall data to support registration and labeling claims.

Cognitive science indicates that memory processes are active and that complex processes control what information is stored in memory and what information can be retrieved (Kahneman, 1999). Importantly, recall is often an active process of reconstruction, requiring subjects to use a variety of mental rules and shortcuts to generate the needed information to complete PRO measures (Hammersley, 1994; Gorin & Stone, 2001). For example, patients estimate the frequencies of behaviors or symptoms based on their recall of those same recall targets in the recent past (Schwarz & Oyserman, 2001). Patients also use more-available information, such as their current state, to formulate answers. For example, patients' recall of past symptoms is influenced by their symptoms at the time of reporting (Eich, Reeves, Jaeger & Graff-Radford, 1985; Smith & Safer, 1993). The immediate testing environment may also influence recall (Schwarz, 1999). Questions based on recall periods usually require the patient not only recall but also to summarize their experiences and we know that this is not accomplished via arithmetic averaging, but rather via cognitive shortcuts (Stone, Broderick, Shiffman & Schwartz, 2004). One commonly used mental shortcut ("cognitive heuristic") leads patients to base their summaries not on an average symptom level, but rather on the peaks of symptom levels and on symptoms that are proximal to the administration of the PRO measure (known as the peak-end effect; (Redelmeier & Kahneman, 1996)). Another mental shortcut indicates that patients with highly fluctuating symptoms bias their summaries upward (Stone, Schwartz, Broderick & Shiffman, 2005), confounding the meaning of recalled symptoms. Thus, retrospective data can be seriously biased by processes that are inherent in the functioning of human memory.

The biases due to such heuristics and contexts effects are especially salient when patients can not truly access the information requested of them and that is one of the problems with extended recall periods (Menon & Yorkston, 2000). While unusual, highly salient events (e.g., a heart attack) may be highly memorable, most PROs in clinical trials are concerned with more routine, variable, and frequent symptom experiences,

where it is likely that recall periods even as short as hours or days may be problematic, and multiple assessments may be required to cover the period of interest (e.g., a two-week baseline or follow-up period). While the field can not currently prescribe the maximal recall period for various symptoms or experiences, the Guidance recommendation that it be kept as short as possible is consistent with the body of empirical research reviewed above.

Importantly, as the Draft Guidance implies, the biases introduced by retrospection are accentuated when individuals are asked to perform additional complex cognitive operations, such as averaging their experience over time or comparing experiences over two time points (Guyatt, Norman, Juniper & Griffith, 2002; Norman, Stratford & Regehr, 1997). Research has specifically challenged global impressions of change, which require that patients assess their current state, recall their pre-treatment state, and make a valid quantitative comparison of the two. Research on biases in autobiographical memory demonstrates that this arrangements elicits significant biases, particularly in the context of a treatment study (Ross, 1989), and studies specifically evaluating global impressions of change against contemporaneous pre-post data have suggested that global impressions are often biased and inaccurate (Stone, 2005).

In contrast with the scientific literature indicating substantial bias in recalled and summarized experience, the literature indicates that queries about the patient's current state do not invoke these biases. A body of literature suggests that questions about current state (or very recent state) are answered by retrieval from experiential memory, whereas questions about longer periods invoke beliefs and heuristic-driven inferences from semantic memory (Robinson & Clore, 2002a; Robinson & Clore, 2002b).

Thus, we believe **the scientific literature strongly supports FDA's caution against relying on recall and the suggestion to rely instead on patients' reports of their current state**.

Documenting timely, in-protocol completion of diary entries

The Draft Guidance states, "*If a patient diary or some other form of unsupervised data entry is used, the FDA plans to review the protocol to determine what measures are taken to ensure that patients make entries according to the study design and not, for example, just before a clinic visit when their reports will be collected*." (lines 334-337). We commend FDA for recognizing the problem of untimely completion of paper diaries, and setting out a requirement for sponsors to document the timely completion of diaries. The reason for using diaries in clinical research is to avoid the biases that pervade recall data. As the Draft Guidance recognizes elsewhere, recall data can be subject to serious biases (e.g., Hammersley, 1994; Gorin & Stone, 2001), and the Draft Guidance recommends collecting data about patient's momentary state as the best way to avoid such biases (Lines 339-343). Diaries aim to avoid these problems by collecting data from patients close to real time. However, this advantage is negated if patients complete their diaries after-the-fact.

FDA's concern that patients may complete paper diaries out of protocol (e.g., just before a visit) is well-founded. Recent research studies (Stone, Shiffman, Schwartz, Broderick, & Hufford, 2002; Broderick, Schwartz, Shiffman, Hufford, & Stone, 2003; Rogers et al, 2006) have demonstrated that <u>most</u> paper diaries are not completed in a timely way that conforms to the protocol and to FDA's guidelines for clinical data (US FDA, 2004). Indeed, the data show that paper diary entries were not only often back-filled, but also demonstrated that forward-filling (i.e., completing a diary entry <u>before</u> its designated time) was common. Obviously, this leads to invalid data and can threaten the validity of a clinical trial. These findings are consistent with a substantial literature in which patients are asked to make paper diary entries of particular events (e.g., taking a medication), which also shows that patients complete such written diaries after-the-fact, and falsify the timing of the entry to give the appearance of timely completion (Osterberg & Blaschke, 2005). Based on these empirical findings, the Agency is well justified in requiring that sponsors who use diaries clearly document timely completion of the diaries, in order to preserve the validity of the data. Modern methods such as electronic diaries can provide for time-tagging to document the time of diary completion, as well as facilitate timely diary completion, and are capable of achieving very high rates of compliance with diary protocols (Hufford & Shiffman, 2003)). **Accordingly, we support the FDA's requirements that protocols using unsupervised PRO assessments such as diaries document timely and in-protocol completion of the assessments.**

Assessing actual behavior rather than patient beliefs, where possible

We support the Draft Guidance's determination that actual behaviors should be measured in PRO assessments, rather than perceived ability to perform behaviors in hypothetical situations (Lines 302-308). This arises most specifically in the context of assessing physical capabilities, but may arise in other contexts as well. As noted in the Draft Guidance, a patient may believe that they can enact a behavior, yet been unable to actually do so. If one is interested in patients' perceived ability, such hypothetical reports may be useful. However, where actual functioning is the issue, hypothetical measures of ability are an inappropriate target for PRO assessment in clinical trials. If behavior is a study endpoint, it is important to measure what patients actually do in real-world settings, not what they believe they can do.

Implied standards and burden for migrating assessments from print to text-based electronic format should be based on scientific evidence.

The Draft Guidance (Lines 178-179) states "*When considering an instrument that has been modified from the original, the FDA generally plans to evaluate the modified instrument just as it would a new one*." This might be read to imply that any change made in the instrument would require *do novo* validation. Comments elsewhere in the Draft Guidance make clear that FDA does not intend to set up such an absolutist requirement for revalidation. Notably, Lines 582-583 state that "*The extent of additional validation recommended depends on the type of modification made*." This implies a more flexible standard that recognizes different degrees of modification., some of which might require little or no re-validation. Presumably, the FDA would determine the extent of revalidation required based on sound scientific principles, including the existing evidence base.

Subsequently, in describing modifications that would lead FDA to consider the modified assessment as being recommended for re-validation (lines 590-593), the Draft Guidance lists the "migration" of a PRO instrument from paper-and-pencil administration to administration on a computer (Lines 636-638). In the specific case where a paper-and-pencil assessment is adapted to be administered on a computer screen, we believe the scientific evidence has already clearly demonstrated equivalence and that, accordingly, no further validation should be necessary.

Specifically, we have reviewed an extensive literature on this question, comprising 58 studies that directly compared paper-and-pencil and computer-administered versions of a wide range of instruments. (Citations are given in an Appendix to this comment.) The studies included administration of instruments on desk-top computers, via the web, on touch screens, and on PDAs, and studies where items were displayed on screen one item at a time (which is not characteristic of paper-and-pencil administrations). The review was limited to computer administrations that display the assessment text on screen and accept physical responses such as pointing or clicking. Studies using other "electronic" media, such as IVRS, were not included, as IVRS is deemed to introduce additional changes in item presentation and response modality (see below).

Among these 58 studies, over 95% were deemed to demonstrate equivalence. A total of 41 unique studies provided sufficient detail to allow quantitative summary and comparison across 235 scales (many studies administered and analyzed multiple scales in both modalities). Among 197 direct comparisons for the equivalence of means, 91% demonstrated no significant difference between means. To quantify the magnitude of the observed differences between the two modes of administration, we scaled the observed difference as a percentage of the range of the scale in question (i.e., if the difference was 3 points on a 1-100 scale, it was listed as 3%). The mean difference between paper and computer administrations averaged -0.3% of the scale range, and was within ±5% of the scale range in more than 90% of all comparisons. Among 180 reported correlation coefficients between paper and computer administration (typically reported as the intraclass correlation coefficient), the average was 0.87; 95% of correlations were at least 0.70, the standard for comparison recommended by Hahn et al (2006, in preparation).

In evaluating the correlations between paper and electronic administrations, it is essential to also consider the test-retest correlation for repeated administrations of the paper instrument. Some variation is observed even in within-mode paper-paper test-retest: the cross-mode correlation of paper and computer test-retest cannot possibly be higher than that. In 20 comparisons that evaluated both, the average cross-mode paper-to-computer correlation was identical to the within-mode correlation for paper measures (both = 0.87). This suggests that the computer-administered versions were completely equivalent to the paper versions. Importantly, results were similar across computer platforms used to administer PRO assessments using text displays of the scale content. Thus, extensive evidence indicates that paper- and computer-administered PROs are psychometrically equivalent.

Because there is already such extensive and positive evidence documenting the equivalence of paper and computer-based assessments, we believe it would be scientifically unjustified to require revalidation when instruments are ported from one administration format to the other. In light of the conclusions from 58 studies and nearly 200 comparisons, requiring further validation would be wasteful and unreasonably burden the use of electronic administration of PROs, which otherwise has substantial advantages and is preferred by patients (e.g., Kvien et al., 2005; Ryan et al., 2002). **Accordingly, migration of a written assessment to computer administration should require only small-scale cognitive debriefing to make sure that the item content has been reasonably formatted for display on the computer platform.**

Other uses of electronic technology in assessment and other adaptations

We also note that the Draft Guidance (lines 636-638) groups administration of a PRO assessment by computer in the same list of examples as administration by IVRS and computer-adaptive testing. We believe that these are very different technologies that raise different issues, and should occasion different degrees of concern and different requirements for re-validation. When a paper-and-pencil instrument is administered on a computer, using a text display to display items and a physical gesture (mouse click or pointing with a stylus) to indicate responses, there is very little change in the subjects' experience and task: The subject can read the item, can see all of the response options at one time, and can review the item and response options as much as necessary. Thus, the subject task is very similar, which accounts for the near-perfect equivalence of the two modes of administration observed in our review.

In contrast, an IVRS system presents subjects with quite a different task and experience. The item text is presented aurally, which requires focused attention and serial processing; the first part of the question may be forgotten even as the last part is being played for the subject. Serial presentation can be an even greater challenge for response options, because the subject must try and remember all response options as they are played back in order to give a valid response. The challenges of aural presentation likely grow as the length of the item and the number of response options grows, in ways that have not been fully explored. Finally, whereas responses on paper (or a computer screen) are typically presented in a meaningful order that helps subjects place themselves in the response set, responding on a telephone keypad may disrupt this ordered physical representation of responses. Thus, transforming a PRO assessment from written text to IVRS is a much bigger change than moving the text from paper to a computer screen. Accordingly, the evidence, cited above, for equivalence of

text-based computer administration cannot be generalized to IVRS, which will need to be evaluated on the basis of IVRS-specific studies, which are beginning to be conducted.

The Draft Guidance also mentions computer-adaptive testing as a change in form of administration. However, computer-adaptive testing involves much more than a change in format – it fundamentally changes what content is assessed and disrupts the usual arrangement in which all subjects respond to the same assessment items. Thus, adaptive testing represents a fundamental change both in the concrete practice of assessment and in the underlying psychometric and statistical theory. Much work is being undertaken to establish the validity of adaptive testing (and item response theory), and the FDA will undoubtedly need to evaluate this evidence as the field moves forward, but the issues of adaptive testing are quite different from and more significant than those encountered when a paper-and-pencil instrument is simply moved intact to a computer for on-screen administration.

Similarly, the same section of the Draft Guidance refers to adapting an interview-based instrument for self-administration (Lines 634-635) or changing the instructions for the assessment (Lines 639-641). These are substantial changes and transformations of the instruments, and, as we have argued above, not analogous to the much simpler adaptation of moving text from paper to screen, whose equivalence has been well-established in dozens of studies.

Thus, we believe **the Draft Guidance should distinguish simple "migration" of an instrument from paper-and-pencil to computer administration from the more complex issues raised by IVRS and adaptive testing. Given the substantial and positive evidence that computer-based administration is equivalent to paper-based administration, porting an instrument from one form to the other should require no revalidation beyond small-scale cognitive testing to ensure that the content was presented and formatted reasonably on the computer display.**

Specific concerns when using electronic PRO instruments

The Draft Guidance addresses an issue that falls outside the scientific validation of PRO assessments and addresses the proper treatment of PRO data that are collected electronically directly from the patient. We commend the FDA for reinforcing the need to ensure that electronic methods used to capture PRO data meet predicate rules, 21 CFR 11, and the Guidance on use of computerized systems in clinical trials. In particular, we support the FDA's specific comments that electronic systems need to have secure and limited access, protect against viewing unblinded data, facilitate FDA review and as required copying of the data, and ensure that data are backed up to multiple locations. We concur that these are among the key components in ensuring the integrity of clinical data.

We appreciate the FDA's need to delineate concerns regarding investigator's responsibility for maintaining case histories. We believe, however, that more explicit expression of FDA's position is needed regarding the following statement from the Guidance (lines 827-835): "*The investigator's responsibility to control, access, and maintain source documentation can be satisfied easily when paper PRO instruments are used, because the subject usually returns the diary to the investigator who either retains the original or a certified copy as part of the case history. The use of electronic PRO instruments, however, may pose a problem if direct control over source data is maintained by the sponsor or the contract research organization and not by the clinical investigator. The FDA considers the investigator to have met his or her responsibility when the investigator retains the ability to control and provide access to the records that serve as the electronic source documentation for the purpose of an FDA inspection*."

We concur with this statement, as we believe it properly emphasizes the investigator's "ability to *control* and *provide access to*" such ePRO records (lines 833-834, our emphasis), as distinct from being in physical possession of the records or the device on which the records reside. Whereas in the paper PRO environment one can use location to demonstrate control, in an electronic environment control is distinct from location or physical possession. In handling electronic data, many logical or electronic security measures can be used to establish control by the investigator. These include limited access and data editing rights, audit trails for changing and even viewing data, and clear procedures. In the electronic realm, the investigator is able to exercise control over the data without the data being physically located at the investigative site. (Indeed, in a modern networked environment, physical possession of the storage device such as a storage computer would not necessarily imply exclusive control without such security and access controls and procedures.) We read the Draft Guidance to imply that FDA is in agreement that control, rather than location, is key to the investigator's exercise of his or her responsibilities. We recommend that the Guidance state so explicitly. Such explicit clarification would facilitate industry understanding and in turn promote best practices to establish investigator control of data when it is not at the same physical location as the investigator. With respect to specific text around this distinction, we recommend the following: "In instances where electronic PRO instrument data do not reside local to the investigative site, it is recommended that the investigator be able to produce clear documentation regarding how control over the data was established and maintained."

Concerns about missing data

We commend the draft guidance for noting the challenges posed by missing data (Lines 755-761). Missing data can occur at two levels: (a) items within an assessment can be missing and (b) entire assessments can be missing. The guidance notes a number of strategies for dealing with these scenarios, including taking an average of scores within an assessment when less than 25 percent of data are missing and using the last observation carried forward technique for missing assessments, among others. As noted in the guidance, all of these strategies are imperfect solutions.

Clearly, *prevention* of missing data is preferable to managing missing data after the fact. Therefore, we think it is important to note methods that can reduce the prevalence of missing data. Electronic data capture can address both sources of missing data. Electronic assessments can ensure presentation of all relevant items and require a participant to respond to an item before progressing on to the next item, therefore eliminating the possibility of missing data within an assessment. Although electronic data capture can not prevent study participants from missing clinic appointments or dropping out of a study, it is useful for diary studies, where assessments are administered to patients outside of the clinic. Electronic platforms (e.g., PDAs, IVRS) can prompt patients to make required diary entries and increase compliance with protocols, resulting in fewer missing assessments. Using strict standards for computing compliance (# of completed entries / # called for by protocol), as suggested in reporting standards for diary data (Stone & Shiffman, 2002), electronic diaries can achieve 90% compliance with timely completion of diary entries, even in protocols that call for many assessments per day (Hufford & Shiffman, 2003; Kamarck et al, 1997; Stone et al, 2003). This minimizes concern about missing data and the representativeness of the data that are available for analysis. Protocols with many observations per patient also lend themselves to analyses that can better handle missing data with a minimum of bias and without the need for casewise deletion or complex data imputation (Walls & Schafer, 2006).

In sum, **we agree that missing data pose a significant challenge to the integrity of a clinical trial and urge the FDA to note methods that may decrease the likelihood of missing data and mitigate its effects on analysis**.

Respondent burden and automated skip patterns

The Draft Guidance (Lines 432-458) notes that respondent burden is an important issue to consider when PRO assessments are administered. We agree. In particular, PRO assessments that require the participant to navigate through complex skip patterns may lead to noncompliance and/or invalid responding. This may not only influence responses covered by the skip pattern, but, by making the assessment burdensome and confusing, may undermine the entire assessment. Missing data resulting from overly burdensome assessments present a challenge in data analysis and interpretation of results. Accordingly, assessments that limit the complexity and frequency of skip patterns may result in higher quality data. When complex item branching is required in order to address endpoints of interest, automated skip patterns may reduce respondent burden. Automated skip patterns are possible on electronic platforms, in which an assessment program can automatically determine appropriate skip patterns, without requiring effort by the participant.

Pooling of data from different assessment modalities.

The Draft Guidance states (Lines 322-324) that "The FDA intends to review the comparability of data obtained when using multiple modes of administration to determine whether pooling of results from the multiple modes is appropriate." We agree that pooling of data from multiple modes within a single trial may represent a particularly high bar for comparability of measures obtained by different modalities. Even in this instance, a showing of comparability may allow pooling, especially if the statistical analysis models any differences in measurement modality (e.g., by inclusion of a dummy variable for modality). However, the more typical case is one where each trial in a program uses a single measurement modality, even if mode of measurement varies across studies (e.g., Phase II using paper-based assessments, Phase III using electronic). In this instance, pooling of data across modalities is not necessary, and there would be no need for special review beyond the considerations mentioned elsewhere in the Guidance. We believe **the Guidance should be clarified to distinguish between the higher bar set for <u>pooling</u> data across modalities within a trial and the lower bar for changing modalities between studies without pooling data.**

Assessment frequency and timing

While most of the Draft Guidance deals with issues involved in validating a particular assessment tool such as a questionnaire, the Draft Guidance also properly addresses the frequency and timing of assessment in clinical trials, noting that "The frequency of PRO assessment depends on the natural history of the disease and the nature of the treatment. Some diseases, conditions, or study designs may necessitate more than one baseline assessment and several PRO assessments during treatment" (Lines 770-776). We support this essential point.

The frequency and scheduling of assessments should be based on an understanding of how the assessment target varies over time. This is particularly important when assessments focus on symptoms, which can vary substantially over time. Designs that rely on one or a few pre-post assessments are incapable of documenting the course of symptoms (e.g., the rate of recovery or possible rebound). More frequent assessments allow the sponsor, agency, physicians, and patients to better understand the trajectory of improvement that is to be expected from a particular treatment (e.g., Stone 2005; Gendreau et al., 2003). Assessment frequency and timing also needs to be sensitive to variation of symptoms even within a day, whether due to diurnal effects or to idiosyncratic volatility in symptoms. Because physicians and scientists typically only encounter patients on a small number of occasions, and in restricted settings, investigators often underestimate the variability of symptoms and assume an unrealistic stability. For example, Stone et al. (2004) showed that pain levels experienced by "chronic" pain patients vary considerably within and across days. In some syndromes, there are known diurnal effects: e.g., in melancholic depression, depressive mood tends to be worst in the morning (APA, 1994; von Zerssen et al., 1987), and pain in rheumatoid arthritis patients tends to peak in the morning (Stone et al., 1997). In these instances, even a daily diary, typically completed in the evening, would not only fail to capture this variability, but could yield highly biased results (e.g., underestimating depression symptoms and pain.). Thus, it is essential for studies of symptoms to consider symptom variability and to plan for frequent assessment to capture this variability. Diaries are a suitable method for capturing such data, and electronic diaries have demonstrated the ability to capture multiple data points within a day, and to characterize variations in symptoms across days and as well as diurnal patterns within a day (Stone et al., 1996) .

Thus, the FDA's emphasis in the Draft Guidance on fitting assessment frequency and timing to the expected natural history of the target symptoms is well-taken. The FDA should encourage the use frequent assessments with electronic diaries when it is useful to capture variations in symptoms.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th ed.).* Washington, DC: American Psychiatric Association.

Broderick JE, Schwartz JE, Shiffman S, Hufford MR, Stone AA. Signaling does not adequately improve diary compliance. *Annals of Behavioral-Medicine*. Vol 26(2) Oct 2003, 139-148.

Broderick, J., Schwartz, J., Shiffman, S., Hufford, M., & Stone, A. (2003). Signaling does not insure diary compliance. *Annals of Behavioral Medicine, 26,* 295-305.

Eich, E., Reeves, J., Jaeger, B., & Graff-Radford, S. (1985). Memory for pain: Relation between past and present pain intensity. *Pain, 223,* 375-379.

Gorin, A. A., & Stone, A. A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. In A. Baum, T. Revenson & J. Singer (Eds.), *Handbook of Health Psychology.* (pp. 405-413). NJ: Erlbaum.

Guyatt, G., Norman, G., Juniper, E., & Griffith, L. (2002). A critical look at transition ratings. *Journal of Clinical Epidemiology, 55,* 900-908.

Hahn, E.A., Cella, D., Chassany, O., Fairclough, D., Wong, G., & Hays, R.D. (2006). *A guide for clinicians to compare the precision of health-related quality of life data relative to other clinical measures*. Manuscript in preparation.

Hammersley, R. (1994). A digest of memory phenomena for addiction research. *Addiction, 89,* 283–293.

Hufford, MR & Shiffman, S (2003). Patient-Reported Outcomes: Assessment Methods. *Disease Management & Health Outcomes*, *11*, 77-86.

Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener & N. Schwarz (Eds.), *Well-being: The foundations of a hedonic psychology* (pp. 3-25). New York: Russell Sage Foundation.

Kamarck, T.W., Shiffman, S., Smithline, L., Thompson, H., Goodie, J., Paty, J.A., Gnys, M., & Kassel, J.D. (1997). The effects of psychosocial influences on ambulatory blood pressure: Contrasting different measurement and data analytic strategies. *Psychophysiology*, 34, S6.

Kvien TK, Mowinckel P, Heiberg T, Dammann KL, Dale O, Aanerud GJ, Alme TN, Uhlig T. (2005). Performance of health status measures with a pen based personal digital assistant. *Annals of Rheumatoid Diseases*.;64(10):1480-4.

Menon, G. & Yorkston, E.A. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgements. In Stone, A.A., Turkkan, J.S., Bachrach, C.A., Jobe, J.B., Kurtzman, H.S., & Cain, V.S. (Eds.), *The Science of Self-Report: Implications for Research and Practice*. Erlbaum: Mahwah, NJ.

Menon, G., & Yorkston, E. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgements. In A. Stone, J. Turkkan, C. Bachrach, J. Jobe, H. Kurtzman & V. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 63-79). Mahwah, N.J.: Lawrence Erlbaum Associates.

Norman, G., Stratford, P., & Regehr, G. (1997). Methodological problems in retrospective computation of responsiveness to change: The lessons of Cronbach. *Journal of Clinical Epidemiology, 50,* 869-879.

Osterberg, L, & Blaschke, T (2005). Adherence to medication. *New England Journal of Medicine,* 353, 487-497

Redelmeier, D., & Kahneman, D. (1996). Patients' memories of pain medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66,* 3-8.

Robinson, M., & Clore, G. (2002a). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128,* 934-960.

Robinson, M., & Clore, G. (2002b). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology, 83,* 198-215.

Rogers, J.M., Katayama, M., Hufford, M., Campbell, E.A., Kling, T., and Kaladas, J.A. (2006). *Electronic diary improves the quality of data collection over conventional paper diary in patients with symptomatic copd*. Annual Meeting of the American Thoracic Society, San Diego, California, May

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96,* 341-357.

Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. *Quality of Life Research,* Vol 11(1) 2002, 19-26.

Schwarz, N. (1999). Self-reports. How the questions shape the answers. *American Psychologist, 54,* 93-105.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication and questionnaire construction. *American Journal of Evaluation, 22,* 127-160.

Smith, W., & Safer, M. (1993). Effects of present pain level on recall of chronic pain and medication use. *Pain, 55,* 355-361.

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient non-compliance with paper diaries. *British Medical Journal, 324,* 1193–1194.

Stone, A., Broderick, J., Porter, L., & Kaell, A. (1997). The experience of rheumatoid arthritis pain and fatigue: Examining momentary reports and correlates over one week. *Arthritis Care and Research, 10,* 185-193.

Stone, A., Broderick, J., Shiffman, S., & Schwartz, J. (2004). Understanding recall of weekly pain from a momentary assessment perspective: Absolute accuracy, between- and within-person consistency, and judged change in weekly pain. *Pain, 107,* 61-69.

Stone, A., Broderick, J., Shiffman, S., & Schwartz, J. (2004). Understanding recall of weekly pain from a momentary assessment perspective: Absolute accuracy, between- and within-person consistency, and judged change in weekly pain. *Pain, 107,* 61-69.

Stone, A., Schwartz, J., Broderick, J., & Shiffman, S. (2005). Variability of momentary pain predicts recall of weekly pain: A consequence of the peak (or salience) memory heuristic. *Personality and Social Psychology Bulletin, 31,* 1340-1346.

Stone, A., Shiffman, S., Schwartz, J., Broderick, J., & Hufford, M. (2002). Patient non-compliance with paper diaries. *British Medical Journal, 324,* 1193-1194.

Stone, A., Shiffman, S., Schwartz, J., Hufford, M., & Broderick, J. (2003). Patient compliance with paper and electronic diaries. *Controlled Clinical Trials, 24,* 182-199.

Stone, A., Smyth, J., Pickering, T., & Schwartz, J. (1996). Daily mood variability: Form of diurnal patterns and determinants of diurnal patterns. *Journal of Applied Social Psychology, 26,* 1286-1305.

Stone, A.A,. (2005). *Real-time ePRO data collected using eDiaries: Comparisons to recall-based measures of change.* Presentation at a DIA conference, Benefits and challenges with epro: Outcomes and endpoints as we emerge from the long age of paper methods, Arlington, VA, April.

Stone, A.A. & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, 24, 236-43.

Stone, A.A., Broderick, J.E., Schwartz, J.E., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: Reactivity, compliance, and patient satisfaction. *Pain, 104,* 343-351

Unites States Food and Drug Administration. (2004). Guidance for industry: Computerized systems used in clinical trials, *Federal Register*, September 2004.

von Zerssen, D. P. Doerr, H. M. Emrich, R. Lund & K. M. Pirke Diurnal variation of mood and the cortisol rhythm in depression and normal states of mind. *European Archives of Psychiatry and Clinical Neuroscience*, 1987, 237, 36-45.

Walls, T.A. & Schafer, J.S. (Eds.) (2006). *Models for intensive longitudinal data.* New York: Oxford University Press.

Willke RJ, Burke LB, Erickson P. (2004). Measuring treatment impact: A review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Applied Clinical Trials*, 25(6):535-52.

Respectfully submitted,

invivodata, inc.
2100 Wharton St (505)
Pittsburgh, PA 15203