ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Systematic Literature Review

Evaluating Recall Periods for Patient-Reported Outcome Measures: A Systematic Review of Quantitative Methods

Cara Arizmendi, PhD, Suwei Wang, PhD, Samantha Kaplan, PhD, Kevin Weinfurt, PhD

ABSTRACT

Objectives: The current guidance for selection of recall periods recommends considering the design of the study, nature of the condition, patient's burden and ability to recall, and intent of the outcome measure. Empirical study of the accuracy of recall periods is recommended; however, there is not consensus on how to quantitatively evaluate the consistency of results from patient-reported outcome measures (PROMs) with different recall periods. We conducted a systematic review to describe quantitative methods for evaluating results obtained from PROMs with differing recall periods to lay the groundwork for establishing consensus.

Methods: We searched MEDLINE, Embase, Scopus, and American Psychological Association PsycINFO for studies where participants are given the same health-related measure (eg, quality of life, well-being, functioning, and pain) with differing recall periods.

Results: A total of 7174 abstracts were screened. The 30 included studies reflected a wide range of domains, including pain, fatigue, and sexual behavior and function. The recall periods ranged from momentary to 6 months. The analytic approaches varied, including different methods for assessing relative agreement, absolute agreement, and for assessing combined relative and absolute agreement.

Conclusions: We found variability in how PROM recall periods were evaluated, suggesting an opportunity for greater consensus on methodological approach. As a starting point, we provide recommendations for which methods are preferred for which contexts.

Keywords: patient-reported outcome measures, quantitative methods, recall periods, systematic review.

VALUE HEALTH. 2024; 27(4):518–526

Introduction

The International Society for Pharmacoeconomics and Outcomes Research and the US Food and Drug Administration (FDA) recommend that researchers carefully consider the selection of recall periods in patient-reported outcome measures (PROMs).[1-3] Current guidance recommends considering the study design, the nature of the condition, the patient's burden and ability to recall, and the concept of interest.[4,5] Empirical studies are suggested as a source of evidence for recall accuracy[3]; however, there is no consensus on which statistical methods should be used to empirically evaluate the recall accuracy of PROMs with different recall periods. Understanding the current state of statistical methods used in empirical evaluations of recall accuracy would lay the groundwork for choosing appropriate recall periods based on empirical evidence.

Importance of Methods for Evaluation of Recall Periods

Because accurate recall of patient experiences is known to degrade over time,[4] it would be ideal to ask patients to report symptoms, feelings, and functions at the time they occur. However, this is not always feasible because of the inability to know when such experiences will occur and the patient burden that constant monitoring would incur. This shortcoming suggests the need to administer PROMs with optimal recall periods so that patients' reporting is well timed and accurate.

Based on current recommendations, there is no one-size-fits-all approach for the selection of recall period; however, factors that are important to the selection of the recall period include the design of the trial, the expected pattern of change, potential issues with adherence to assessment completion, the PROM, the patient burden, the patients' ability to recall, and any impact specific to the patient population and disease state.[1,2,4,5] Any potential interactions of these factors should also be considered.'

The FDA and review articles make suggestions for recall period selection based on the above factors.[2-5] When the FDA evaluates PRO-based labeling claims, they will investigate whether an effort was made to "ensure that patients understood the instrument recall period."[2] FDA guidance also recommends that recall periods correspond to the schedule of assessment and that a shorter recall

period is used when appropriate.[2] Additionally, the FDA suggests that when there is doubt about the accuracy of a proposed recall period, sponsors provide evidence from existing literature and/or new empirical studies.[3]

When it is not always known what recall period will be most appropriate for a given population or disease, as suggested by the FDA, it can be beneficial to empirically evaluate whether one recall period (eg, 7 days) obtains similar results as another recall period (eg, 1 day). For example, if, indeed, the 7-day recall period obtains approximately the same results as the 1-day recall period, then choosing the 7-day recall period would reduce patient burden. However, guidelines for how to compare the similarity of PROM results that use different recall periods are currently lacking. As a first step in creating these guidelines, we conducted a systematic review to assess what metrics researchers currently use to empirically evaluate similarity in recall periods.

Specific Aims

Our study reviewed all empirical studies that quantitatively evaluated the difference in results obtained from PROMs using different recall periods. Our specific aims were to (1) describe current methods for empirical evaluation of the similarity of results obtained from different recall periods and (2) offer preliminary recommendations for empirical comparison of the similarity of results obtained from different recall periods.

It should be noted that the original goal of this project was to summarize and report findings about the accuracy of different recall periods using meta-analysis from the studies we found; however, the heterogeneity of the findings made drawing conclusions about the results untenable. Instead, we focused on reporting the methods used in these studies in the hopes of creating more consistency in future recall period studies that will benefit future meta-analytic work.

Methods

We searched for studies that compared results obtained from different recall periods from the same PROM, in which the only difference between PROMs was the wording of the recall period. Health-adjacent reports that are not measures of feeling or function (eg, environmental exposure and religious attendance) were not included. We also did not include studies that looked at patients' change in perception after an event (eg, recall of initial traumatic brain injury symptoms 1-month post-injury vs 3 months post-injury). Finally, we did not include studies in which accuracy was assessed in comparison with "objective" measures (eg, recall of exercise vs pedometer data) or in which accuracy was compared between PROs and observer/clinician-reported outcomes. Although we believe that these may be important methods for assessing accuracy, they are outside the scope of the current article.

Search

We searched for studies in which participants are given the same health-related measure (eg, quality of life, well-being, functioning, and pain) with differing recall periods and in which the results obtained from the differing recall periods are compared. A medical librarian with expertise in systematic searching developed a search utilizing a mix of subject headings and keywords to represent the concepts of recall, survey, symptoms, quantitative analysis, and time factors. The databases Ovid MEDLINE, Embase via Elsevier, Scopus via Elsevier, and American Psychological Association PsycINFO were searched from inception to November 8, 2021. To ensure that results were up-to-date, an additional search from 8 November 2021 to 20 July 2023 was conducted but did not yield any additional studies for inclusion. All search results were compiled in EndNote and imported into Covidence for deduplication and screening. Search strategies are available in Appendix A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2024.01.01.

Screening

A total of 7174 abstracts were found and screened independently by the first and second authors. If there was a conflict in the recommendation to include or exclude the study, the 2 screeners consulted with the last author to reach a consensus. Subsequently, 124 studies received an independent, full-text review by the first and second author. Again, consensus was reached based on discussion between the authors. Of the 124 full-text studies screened, 94 were excluded from extraction for reasons including that the study did not make a quantitative comparison of different recall periods in PROMs (n = 48); different items were used in different recall periods (n = 17); participants were asked to recall an event from the far past (n = 16), measures of feeling or function were not used (n = 8); an objective versus a subjective comparison was made (n = 3); and prevalence observed in one recall period was compared with another (n = 2). Ultimately, 30 studies were included for data extraction. Figure 1 shows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses study flowchart.

Data Extraction

We extracted the following general information from the 30 studies included in our review: title, year, health domains, PROMs, sample size, participant age range, participant health conditions, and the length of the recall periods included in the study. We also extracted the comparison metric (eg, raw difference, Pearson correlation coefficient, differential item functioning); which methods were used to aggregate the shorter recall period (eg, mean, maximum, and last day); whether the comparison was made at the item response, subdomain, or total score level; whether the article looked at the influence of patient-level variability on similarity results; and whether the article looked at the impact of level of feeling or function on similarity results. For each comparison metric used in a study, we extracted all statistical indices the investigators used for the comparison (eg, Pearson correlation and paired *t* test) and the method used for inference (eg, confidence interval, *P* value).

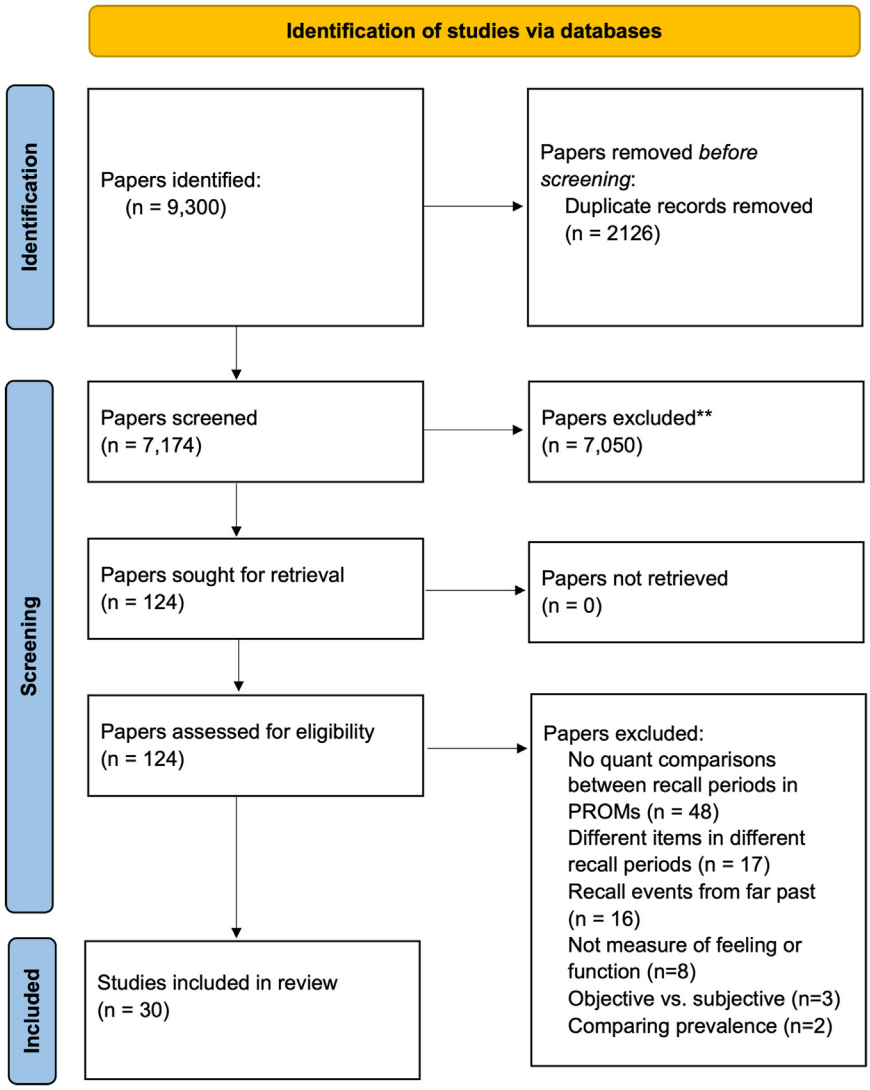
Analysis

We report counts for the domains, PROMs, and patient conditions under study along with whether comparisons were made at the item response, subdomain, or total score level. Additionally, we report the comparison metric used to compare recall periods. For each type of comparison metric, we provide a description of the number of studies using the comparison metric. Finally, during the course of extraction, we noticed that several studies looked at potential moderators to the level of agreement between recall periods. Thus, we also report the moderators that the studies examined and the number of studies exploring each of these potential moderators.

Results

The 30 studies that were included reflect a wide range of domains, PROMs, patient diseases and conditions, recall periods, and methods for assessing similarity between results from different recall periods (Table 1).[6–35]

Figure 1. PRISMA study flowchart.



PRISMA indicates Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Domains, PROMs, and Patient Conditions

The domain most frequently assessed was pain or pain interference (n = 11). This was followed by fatigue (n = 5), sexual behavior and function (n = 4), and mood-related measures (n = 3). There was even greater variability in the PROMs being assessed in each study. In fact, many studies (n = 13) assessed a set of questions that were unique to the study (ie, not from an established PROM or item bank). The most commonly used PROMs were the Brief Fatigue Inventory (n = 3) and the Brief Pain Inventory (n = 3). Similarly, there was variability in the health conditions under study. Approximately one-quarter (n = 8) of the studies had participants from the general population. The only conditions covered by >2 articles were rheumatological diseases (n = 3).

Length of Recall Periods

More than one-third of studies (n = 11) made comparisons between daily and weekly recall periods, and around 16% (n = 5) made comparisons between daily and biweekly recall periods. Twenty percent (n = 6) made comparisons between daily and monthly recall periods. Ten percent (n = 3) made comparisons between momentary assessments and daily assessments, and 13% (n = 4) made comparisons between momentary and weekly recall periods. The remaining comparisons assessed different combinations that included twice daily, twice weekly, 1 week, 2 weeks, 3 weeks, 4 weeks, 1 month, 3 months, and 6 months.

Level of Comparison

Half the studies (n = 15) made comparisons between recall periods only at the item response level, 7 studies only at the subdomain level, and 1 study only at the total score level. Three studies made comparisons between recall periods at both the item response and the subdomain level, 3 at both the item response and total score level, and 1 at both the subdomain and total score level. No studies made comparisons between recall periods at all 3 levels.

Metrics Used to Assess Similarity

The analytic approaches for assessing similarity varied, including different approaches for describing relative and/or

Table 1. Summary of results.

Study N participants	Domains	PROMs	Patient condition	Recall periods	Level of comparison	Metric	Aggregation of shorter recall period	Level of feeling/ function	Patient-level variability	Aggregation method
Bennett et al,[6] 2010 N = 38	Cystic fibrosis respiratory symptoms	CFRSD	Cystic Fibrosis	1 day 1 week	Subdomain	CCC; Pearson and/or Spearman Correlation; Difference (non-standardized)*,†; Difference (t-statistic)*,†	Mean, Median, Mode, Max, Min, First, Penultimate	No	No	Yes
Bennett et al,[7] 2011 N = 140	Type 2 diabetes symptoms and impacts	Set of questions unique to study	Type II Diabetes	1 day 1 week	Total Score; Item Response	CCC*; Difference (non-standardized)*,§; Pearson and/or Spearman Correlation	Mean, Median, Mode, Max, Min, First, Penultimate, Last	No	Yes	Yes
Bennett et al,[8] 2012 N = 98	COPD symptoms	DPD	COPD	1 day 1 week	Subdomain	Difference (non-standardized)*; CCC; Pearson and/or Spearman Correlation	Mean, Median, Mode, Max, Min, First, Penultimate, Last	Yes	Yes	Yes
Boesen et al,[9] 2020 N = 122	Thyroid-related quality of life	ThyPro	Thyrotoxicosis	Momentary 1 week 4 weeks	Subdomain	Pearson and/or Spearman Correlation†; Difference (non-standardized)*	Mean	No	No	No
Brauer et al,[10] 2003 N = 119	Pain and pain interference	PRIM Questionnaire	Other: musculoskeletal disorders	1 week 3 months	Item Response	Proportion of participants with agreement; Weighted Kappa†	Median, Max	Yes	No	Yes
Broderick et al,[11] 2009 N = 105	Fatigue; pain and pain interference	BFI; BPI; McGill Pain Inventory; SF-36; Set of questions unique to study	Osteoarthritis; Other: osteoarthritis, rheumatoid arthritis, lupous, fibromyalgia	Momentary 1 day	Item Response	Difference (t-statistic); Pearson and/or Spearman Correlation‖	Mean, Latent Mean	No	No	Yes
Bushnell et al,[12] 2013 N = 139	Plaque psoriasis symptoms	PSI	Plaque Psoriasis	1 day 1 week	Item Response	Difference (non-standardized) *,†; ICC†	Mean	No	No	No
Coxon,[13] 1999 N = 74	Sexual behavior	Set of questions unique to study	None	1 day 1 month	Item Response; Total Score	Relative difference; Pearson and/or Spearman Correlation	Sum	No	No	No
Dunn et al,[14] 2010 N = 29	Medication use; pain and pain interference; pain behavior or self-care	Set of questions unique to study	Lower Back Pain or Back Pain	1 day 2 weeks	Item Response	Kappa†; Difference (non-standardized)†BA; ICC (2, 1)*	Mean	No	No	No
Flynn et al,[15] 2019 N = 515	Lower urinary tract symptoms	Set of questions unique to study	Lower Urinary Tract Symptoms	1 day 1 week 1 month	Item Response	Pearson and/or Spearman Correlation†; Percent bias	Mean, Max, Last	No	No	Yes
Glick et al,[16] 2013 N = 95	Sexual behavior	Set of questions unique to study	None	Twice weekly 1 week 2 weeks 3 months	Item Response	Difference (non-standardized)†; CCC; Kappa	Sum	No	No	No
Jamison et al,[17] 2006 N = 21	Pain and pain interference	Set of questions unique to study	Lower Back Pain or Back Pain	Momentary 1 week	Total Score	Pearson and/or Spearman Correlation*,†,‖	Mean	No	Yes	No
Mark et al,[18] 2017 N = 628	Sexual behavior	Set of questions unique to study	None	1 day 3 months	Item Response; Subdomain	Pearson and/or Spearman Correlation*; Difference (non-standardized)*	Mean, Sum	No	No	No
Martin et al,[19] 2012 N = 220	Pain and pain interference	Set of questions unique to study	Molar Removal	Momentary 1 week	Item Response	ICC†; Difference (antilog)BA	Mean	Yes	No	No

continued on next page

Table 1. Continued

Study N participants	Domains	PROMs	Patient condition	Recall periods	Level of comparison	Metric	Aggregation of shorter recall period	Level of feeling/ function	Patient-level variability	Aggregation method
Matteson et al,[20] 2015 N = 13	menstrual bleeding	Menstrual Bleeding Questionnaire (MBQ)	Abnormal Uterine Bleeding	1 day 1 week 1 month	Subdomain	Pearson and/or Spearman Correlation[†]	Mean	No	No	No
Mendoza et al,[21] 2017 N = 127	Symptomatic adverse events	PRO-CTCAE	Cancer; Other: solid tumor or hematologic malignancy	1 day 1 week 2 weeks 3 weeks 4 weeks	Item Response	Difference (non-standardized)[§]; ICC (3,1)	Mean, Max	No	No	Yes
Mneimne et al,[22] 2019 N = 257	BPD symptoms and triggers; emotions (positive, negative, anger)	Positive and Negative Affect Schedule; Set of questions unique to study	Borderline Personality Disorder; None; Other: current or lifetime disorder other than BPD	1 day 1 week 1 month 6 months	Item Response; Subdomain	Multilevel Model with recall period as predictor (F-statistic)*; Pearson and/or Spearman Correlation*	Mean, Min	Yes	No	No
Pilz et al,[23] 2018 N = 32	Mood/mood-related factors, depression, anxiety	MRI	None	1 day 15 days	Item Response	Proportion of participants with agreement; Difference (non-standardized)[†]; Pearson and/or Spearman Correlation*	Median, Mode	No	No	Yes
Revicki et al,[24] 2009 N = 12	Gastroparesis symptoms	GCSI	Gastroparesis	1 day 2 weeks	Subdomain; Total Score	Pearson and/or Spearman Correlation*	Mean	No	No	No
Ryden et al,[25] 2016 N = 446	Reflux symptoms	RESQ	GERD	Twice daily 1 week	Subdomain	Difference (non-standardized)*; Difference (t-statistic)*; CCC;	Mean, Mode, Last, Sum	No	No	Yes
Schneider et al,[26] 2011 N = 97	Fatigue; pain and pain interference	BFI; BPI	Chronic Rheumotological Disease	Momentary 1 day	Item Response	Hierarchical multiple regression model*,[‡]	Mode	No	No	Yes
Schneider et al,[27] 2013 N = 100	Fatigue; mood/mood-related factors, depression, anxiety; pain and pain interference	PROMIS Depression or Anxiety or Anger; PROMIS Fatigue; PROMIS Pain (and related)	None	1 day 1 week	Item Response	DIF*	Mean	No	No	No
Self et al,[28] 2015 N = 50	Pain and pain interference; stool symptoms	Bristol Stool Form Scale; Set of questions unique to study	IBS	Momentary 2 weeks	Item Response	ICC; Pearson and/or Spearman Correlation*; Difference (non-standardized)[BA]	Max, Sum	No	No	No
Simsek et al,[29] 2008 N = 20	Behcet's disease symptoms	Behcet's disease Current Activity Form; IBDDAM	Behcet's disease	1 day 4 weeks	Item Response	ICC[†]	Sum	No	No	No
Stewart et al,[30] 1999 N = 132	Headache pain and impact	HImQ	Migraine	1 day 3 months	Item Response; Total Score	Pearson and/or Spearman Correlation	Mean, Sum	Yes	No	No
Stone et al,[31] 2004 N = 68	Pain and pain interference	Set of questions unique to study	Chronic Pain	Momentary 1 week	Subdomain	Difference (t-statistic); Pearson and/or Spearman Correlation*; ICC (A,1), ICC (C,1)[†]*	Mean	No	No	No
Stone et al,[32] 2010 N = 106	Fatigue; pain and pain interference	Brief Fatigue Inventory (BFI); Brief Pain Inventory (BPI)	Chronic Rheumatological Disease	Momentary 1 day	Item Response	Z-statistic; Pearson and/or Spearman Correlation	Mean, Max, Min	Yes	No	Yes

continued on next page

Table 1. Continued

Study N participants	Domains	PROMs	Patient condition	Recall periods	Level of comparison	Metric	Aggregation of shorter recall period	Level of feeling/ function	Patient-level variability	Aggregation method
Stone et al,[33] 2016 N = 472	Emotions (positive, negative, anger); fatigue; mood/mood-related factors, depression, anxiety; pain and pain interference; pain behavior or self-care; physical function	PROMIS Depression or Anxiety or Anger; PROMIS Fatigue; PROMIS Pain (and related); PROMIS Physical Function	Cancer; Hernia; None; Osteoarthritis; Premenstrual Syndrome	1 day 1 week	Subdomain	Pearson and/or Spearman Correlation*	Mean	No	No	No
Tran et al,[34] 2013 N = 161	Sexual behavior	Set of questions unique to study	None	1 day 2 weeks	Item Response; Subdomain	Proportion of participants with agreement; Kappa*,†; McNemar's Test; Difference (non-standardized)*,†; Pearson and/or Spearman Correlation*,†	Sum	No	No	No
Wood et al,[35] 2015 N = 32	Symptomatic adverse events	PRO-CTCAE	Hematopoietic Cell Transplantation	1 day 1 week	Item Response	Difference (non-standardized)§; CCC	Max	No	No	Yes

BA indicates Bland-Altman method; BFI, Brief Fatigue Inventory; BPD, borderline personality disorder; BPI, Brief Pain Inventory; CCC, Concordance Correlation Coefficient; CFRSD, Cystic Fibrosis Respiratory Symptom Diary; COPD, Chronic obstructive pulmonary disease; DIF, Differential Item Functioning; DPD, Dyspnea Patient Diary; GCSI, Gastroparesis Cardinal Symptom Index; GERD, gastroesophageal reflux disease; HImQ, Headache Impact Questionnaire; IBDDAM, Iranian BD Dynamic Measure; IBS, irritable bowel syndrome; ICC, Intraclass Correlation; Max, maximum; Min, minimum; MRI, Mood Rhythm Instrument; PRIM, Project on Research and Intervention in Monotonous Work; PRO-CTCAE, Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events; PROM, patient-reported outcome measures; PSI, Plaque Psoriasis Inventory; RESQ, Reflux Symptom Questionnaire.
**P* value reported.
†CI reported.
‡Standard error reported.
§Effect size reported.
||That both between- and within-person correlations were reported.

absolute agreement (between PROM scores from the longer recall period compared with the aggregated, shorter recall period), regression analysis, and differential item functioning (DIF). We describe use of each type of method and the approach to statistical inference taken.

Methods that assessed relative agreement

Some approaches focused on the relative consistency of scores without regard for differences in the absolute scores (eg, scores obtained from shorter and longer recall periods could have a Pearson correlation of 1.0 but still differ in their means). Sixty percent of studies (n = 19) reported a Pearson's and/or Spearman's correlation between results from a longer recall period and shorter recall period. Of these studies, most (n = 17) reported solely between-person correlations. One study reported a between-person correlation looking at the relationship between change in weekly and change in momentary pain.[31] Two studies reported both between- and within-person correlations.[11,17] Additionally, different inferential approaches were used for the correlations reported. Seven studies that reported a correlation provided *P* values for the correlations, 3 provided confidence intervals, 2 provided both *P* values and confidence intervals, and 7 provided no inferential statistic for the correlation.

Methods that assessed absolute agreement

Some approaches focused on differences in raw score values, which translates into a difference in the location (eg, mean or median) of the distribution of scores obtained from longer and shorter recall periods. When such a difference was present, it was often referred to as "bias." Almost two-thirds of studies (n = 19) reported the absolute agreement between different recall periods. A range of methods were used, including the raw difference or antilog of the raw difference (n = 15), the t-statistic or z-statistic (n = 5), the effect size (n = 3), and/or the percent bias (n = 1). Of the studies reporting the raw difference, 4 provided *P* values for the raw difference, 3 provided confidence intervals, 2 provided effect sizes, 3 provided both *P* values and confidence intervals, and 1 provided both *P* values and effect sizes. Additionally, a few studies (n = 3) provided a Bland-Altman plot,[36] reporting the limits of agreement as a complement to the difference value.

Methods that combined relative and absolute agreement

Half the studies (n = 15) reported a statistic that represented a combination of relative and absolute agreement between results from longer and aggregated, shorter recall periods. Of these, 7 reported the intraclass correlation coefficient (ICC). For the ICC, 1 study[14] used the formula for ICC(2,1) per Shrout and Fleiss[37]; 1 study[21] used the formula for ICC(3,1) per Shrout and Fleiss[37]; 1 study[31] used both the formulas for ICC(A,1) and ICC(C,1) per McGraw and Wong[38]; 1 study used a model described as "a single measure, 2-way mixed model with absolute agreement"[28]; and 3 studies did not specify the type of ICC calculated.[12,19,29]

Additionally, 6 studies reported the concordance correlation coefficient, and 4 reported Kappa, one of which specified Kappa as weighted Kappa.[10]

Of the 7 studies reporting the ICC, 1 provided *P* values, 3 provided confidence intervals, 1 provided both *P* values and confidence intervals, and the remaining 2 studies provided no inferential statistics. Of those reporting the concordance correlation coefficient, 1 study provided a *P* value, whereas the other 5 studies did not provide an inferential statistic. Finally, of the 4 studies reporting Kappa, 2 provided confidence intervals, 1 provided both *P* values and confidence intervals, and 1 study provided no inferential statistic.

Regression methods

Two studies used regression techniques to assess the similarity between results from shorter and longer recall periods. Both studies used multilevel modeling approaches accounting for within-person variance. One study regressed daily recall on the mean ecological momentary assessment value from that day, controlling for peak and end-of-day values.[26] In this study, the degree of association between recall periods was reflected by the slope parameters and the percent variance explained by mean ecological momentary assessment value. *P* values and standard errors were reported with the slope parameters. The other study used a marginal multilevel model on each item to look at the relationships between results from different recall periods.[22] Differences in results from different recall periods were reflected by the F-statistic with accompanying *P* value.

Other methods

Two other methods did not fit neatly into any of the above categories. One study conducted a DIF analysis, comparing item response theory item parameters between different recall periods. Three studies looked at the proportion of participants that showed some level of agreement in their results from longer and aggregated, shorter recall periods (n = 3). These studies reported the proportion of participants whose responses on items were equal. One study also reported proportions of participants whose responses on items were within 1 or 2 score units when comparing the longer and aggregated, shorter recall period.

Consideration of Moderators to Level of Agreement

In addition to extracting the metrics used to assess similarity, we also reviewed potential moderators to the relationship between results obtained from different recall periods. Of the 30 articles we examined, 6 looked at the impact of the level of feeling or function, 3 looked at the impact of the variability in an individual's responses, and 12 looked at the impact of using different aggregation methods. Methods for aggregating the shorter recall period included calculating the mean (n = 21), the maximum (n = 10), the sum (when comparing reported counts; n = 8), the minimum (n = 5), the median (n = 5), the mode (n = 5), the last day of the shorter recall period (n = 4), the first day of the shorter recall period (n = 3), the next to last day of the shorter recall period (n = 3), and the latent mean (n = 1). Additionally, 4 studies dichotomized the shorter recall period.

Discussion

Overall, this review highlights the substantial heterogeneity in approach, measures, populations, and conclusions in studies examining agreement in responses to PROMs with different recall periods. The studies included a wide range of domains, PROMs, and populations. A variety of methods were used to assess and report the level of agreement, including correlation methods, methods assessing differences, and methods looking at a combination of correlation and difference. Additionally, studies differed in terms of the different combinations of recall periods compared, whether the comparison was made at the item response, subdomain, or total score level, and whether and which potential moderators to the agreement level were considered. Finally, although many studies found some level of agreement between results from different recall periods, there were often caveats made by the researcher as to when that level of agreement holds. This heterogeneity, particularly in the methods used, suggests a lack of agreement on how to best quantitatively characterize the agreement in results between different recall periods; broader discussion is needed to develop a consensus around best approaches. To stimulate discussion, we provide a starting point based on the different methods we found in our literature review.

Preliminary Recommendations

First, when choosing the target of comparison, we recommend reporting results at both the item response level and subdomain/total score level because each might be important for different uses and purposes.

Second, to explore possible reasons for lack of agreement between recall periods, investigators might examine alternative response strategies used by the participants. For example, investigators should examine the relationship between PROM responses to longer recall periods and responses from the last reported day (to test for recency effects) or the maximum/minimum PROM response (to test for a reliance on peaks and valleys).

Third, we recommend that researchers use statistical approaches that provide separate assessments of relative and absolute agreement. The consequences of imperfect recall will depend upon the type of disagreement (relative or absolute) and the intended use of the PROM scores. For example, in a clinical trial designed to compare PROM scores at a fixed follow-up time between 2 randomized groups, the level of the PROM score does not hold as much importance as the relative comparison between the 2 groups. In these cases, a small amount of absolute disagreement (eg, bias) between differing recall periods might not be a concern if the bias can be assumed to be the same in both groups. In contrast, if a PROM score threshold will be used to make decisions about who is eligible to participate in a trial or used as the basis for a categorized endpoint (eg, unimpaired vs impaired), the accuracy of the PROM score holds more importance. In these cases, absolute disagreement, or bias in the level of the PROM score, could result in patients being categorized incorrectly (eg, categorizing a patient as not experiencing impairments when they are experiencing impairments). Additionally, if consistency in the pattern of responses is important, methods such as DIF should be considered.

Fourth, when conducting inferential statistical tests of agreement indices, researchers should use 95% confidence intervals to determine whether the level of agreement in the sample is unlikely to be observed if the true agreement in the population is in an unacceptable range (eg, as in Flynn et al).[14]

Fifth, when interpreting the results from recall studies, investigators should consider the level of feeling or function and variability in feeling or function for each patient.[4,5,22,30] Patients who demonstrate low variability in feeling or function (eg, intense pain every day) may find it easier to recall and summarize past levels of feeling or function than patients with greater levels of variability (eg, pain that changes day-to-day).[6,8] Patient-level variability should especially be noted when assessing correlations or regression coefficients, as smaller variance implies smaller covariance. The level of feeling and function is also important as

more intense feelings may be recalled and summarized differently than less intense feelings (eg, intense pain vs moderate pain).[8]

In addition to deciding how best to conduct research on recall periods, it is important to consider when to collect such evidence. The need for evidence depends on both a priori confidence that recall errors do not overly influence the scores and on the costs/benefits of collecting evidence in the form of a recall study such as those reported here.[39] Regarding a priori confidence, researchers already draw on their experience and that of research participants (eg, from cognitive interviews) to inform their confidence in a recall period. For example, for most settings and health concepts, many are comfortable that a 24-hour recall period is sufficiently accurate and therefore feel no need to verify this empirically. In contrast, there might be less confidence in a PROM with a 30-day recall period, requiring some empirical verification. Regarding the costs/benefits of collecting evidence, recall studies can be burdensome for participants compared with other studies that inform the use of PROMs because participants must complete the same set of questions multiple times on different timescales. Further, it can take time to recruit the sample sizes required to permit sufficiently precise statistical estimates of agreement. These costs and burdens must be weighed relative to the incremental confidence that might be gained in the recall period over and above the a priori level of confidence.

Limitations and Future Directions

This systematic review has a few limitations. One is that we only examined how to assess similarity between different recall periods, but there are many different ways to evaluate whether one recall period is better than another. For example, recall periods may differ in their ability to detect meaningful change. We also restricted studies to PROMs used in clinical trials. There is a broader literature on the ability to recall health-related concepts (eg, hospitalizations) that may provide greater insight into how to make comparisons between information recalled from different lengths of time. Finally, we did not address how these results could complement data from qualitative interviews. Developing a framework for using both qualitative and empirical data could be useful when evaluating the accuracy of recall periods.

Conclusions

The current empirical literature on PROM recall periods is marked by wide heterogeneity of health concepts, measures, and methods. The field would benefit from a more standard approach to assessing recall periods. We hope that our recommendations can serve as a starting point.

Author Disclosures

Links to the disclosure forms provided by the authors are available here.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2024.01.016.

Article and Author Information

Accepted for Publication: January 17, 2024

Published Online: February 28, 2024

doi: https://doi.org/10.1016/j.jval.2024.01.016

Author Affiliations: Population Health Sciences, Duke University School of Medicine, Durham, NC, USA (Arizmendi, Wang, Weinfurt); Medical Center Library, Duke University School of Medicine, Durham, NC, USA (Kaplan).

Correspondence: Cara Arizmendi, PhD, Population Health Sciences, Duke University School of Medicine, Durham, NC 27710, USA. Email: cara.arizmendi@gmail.com

Author Contributions: *Concept and design:* Arizmendi, Wang, Weinfurt,
Acquisition of data: Arizmendi, Wang, Kaplan,
Analysis and interpretation of data: Arizmendi
Drafting of the manuscript: Arizmendi, Wang, Weinfurt
Critical revision of the paper for important intellectual content: Arizmendi, Wang, Weinfurt
Statistical analysis: Arizmendi
Provision of study materials or patients: Kaplan
Obtaining funding: Weinfurt
Administrative, technical, or logistic support: Kaplan
Supervision: Weinfurt

Funding/Support: Cara Arizmendi received full funding from AstraZeneca Pharmaceuticals through the Duke Measurement and Regulatory Science (MaRS) Fellowship. Suwei Wang received full funding from Takeda Pharmaceuticals through the Duke MaRS Fellowship.

Role of the Funder/Sponsor: Sponsors provided funding for postdoctoral fellowship to complete research.

REFERENCES

1. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2–assessing respondent understanding. *Value Health*. 2011;14(8):978–988.
2. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. Food and Drug Administration. https://www.fda.gov/media/77832/download. Accessed July 20, 2023.
3. Patient-focused drug development selecting, developing, or modifying fit-for- purpose clinical outcome assessments: draft guidance. Food and Drug Administration. https://www.fda.gov/media/159500/download. Accessed July 20, 2023.
4. Stull DE, Leidy NK, Parasuraman B, Chassany O. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. *Curr Med Res Opin*. 2009;25(4):929–942.
5. Norquist JM, Girman C, Fehnel S, DeMuro-Mercon C, Santanello N. Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration. *Qual Life Res*. 2012;21(6):1013–1020.
6. Bennett AV, Patrick DL, Lymp JF, Edwards TC, Goss CH. Comparison of 7-day and repeated 24-hour recall of symptoms of cystic fibrosis. *J Cyst Fibros*. 2010;9(6):419–424.
7. Bennett AV, Patrick DL, Bushnell DM, Chiou CF, Diehr P. Comparison of 7-day and repeated 24-h recall of type 2 diabetes. *Qual Life Res*. 2011;20(5):769–777.
8. Bennett AV, Amtmann D, Diehr P, Patrick DL. Comparison of 7-day recall and daily diary reports of COPD symptoms and impacts. *Value Health*. 2012;15(3):466–474.
9. Boesen VB, Feldt-Rasmussen U, Bjorner JB, et al. Shorter recall period for the thyroid-related patient-reported outcome measure ThyPRO did not change the accuracy as evaluated by repeated momentary measurements. *Thyroid*. 2020;30(2):185–191.
10. Brauer C, Thomsen JF, Loft IP, Mikkelsen S. Can we rely on retrospective pain assessments? *Am J Epidemiol*. 2003;157(6):552–557.
11. Broderick JE, Schwartz JE, Schneider S, Stone AA. Can end-of-day reports replace momentary assessment of pain and fatigue? *J Pain*. 2009;10(3):274–281.
12. Bushnell DM, Martin ML, McCarrier K, et al. Validation of the psoriasis symptom inventory (PSI), a patient-reported outcome measure to assess psoriasis symptom severity. *J Dermatol Treat*. 2013;24(5):356–360.
13. Coxon AP. Parallel accounts? Discrepancies between self-report (diary) and recall (questionnaire) measures of the same sexual behaviour. *AIDS Care*. 1999;11(2):221–234.
14. Dunn KM, Jordan KP, Croft PR. Recall of medication use, self-care activities and pain intensity: a comparison of daily diaries and self-report questionnaires among low back pain patients. *Prim Health Care Res Dev*. 2010;11(1):93–102.
15. Flynn KE, Mansfield SA, Smith AR, et al. Can 7 or 30-day recall questions capture self-reported lower urinary tract symptoms accurately? *J Urol*. 2019;202(4):770–778.

16. Glick SN, Winer RL, Golden MR. Web-based sex diaries and young adult men who have sex with men: assessing feasibility, reactivity, and data agreement. *Arch Sex Behav*. 2013;42(7):1327–1335.
17. Jamison RN, Raymond SA, Slawsby EA, McHugo GJ, Baird JC. Pain assessment in patients with low back pain: comparison of weekly recall and momentary electronic data. *J Pain*. 2006;7(3):192–199.
18. Mark KP, Smith RV, Young AM, Crosby R. Comparing 3-month recall to daily reporting of sexual behaviours. *Sex Transm Infect*. 2017;93(3):196–201.
19. Martin WJ, Heymans MW, Skorpil NE, Forouzanfar T. Can a single pain rating replace a multiple pain rating in third molar surgery studies? Analysis of 220 patients. *Int J Oral Maxillofac Surg*. 2012;41(8):1010–1013.
20. Matteson KA, Scott DM, Raker CA, Clark MA. The menstrual bleeding questionnaire: development and validation of a comprehensive patient-reported outcome instrument for heavy menstrual bleeding. *BJOG*. 2015;122(5):681–689.
21. Mendoza TR, Dueck AC, Bennett AV, et al. Evaluation of different recall periods for the US National Cancer Institute's PRO-CTCAE. *Clin Trials*. 2017;14(3):255–263.
22. Mneimne M, Furr RM, Mendrygal D, Law MK, Arnold EM, Fleeson W. Degree of correspondence between retrospective and proximal reports of borderline personality disorder symptoms, symptom triggers, and emotions. *J Pers Disord*. 2021;35(1):1–20.
23. Pilz LK, Carissimi A, Francisco AP, et al. Prospective assessment of daily patterns of mood-related symptoms. *Front Psychiatry*. 2018;9:370.
24. Revicki DA, Camilleri M, Kuo B, et al. Development and content validity of a gastroparesis cardinal symptom index daily diary. *Aliment Pharmacol Ther*. 2009;30(6):670–680.
25. Ryden A, Leavy OC, Halling K, Stone AA. Comparison of Daily versus Weekly Recording of gastroesophageal reflux Disease Symptoms in Patients with a Partial Response to Proton Pump Inhibitor Therapy. *Value Health*. 2016;19(6):829–833.
26. Schneider S, Stone AA, Schwartz JE, Broderick JE. Peak and end effects in patients daily recall of pain and fatigue: a within-subjects analysis. *J Pain*. 2011;12(2):228–235.
27. Schneider S, Choi SW, Junghaenel DU, Schwartz JE, Stone AA. Psychometric characteristics of daily diaries for the Patient-Reported Outcomes Measurement Information System (PROMIS®): a preliminary investigation. *Qual Life Res*. 2013;22(7):1859–1869.
28. Self MM, Williams AE, Czyzewski DI, Weidler EM, Shulman RJ. Agreement between prospective diary data and retrospective questionnaire report of abdominal pain and stooling symptoms in children with irritable bowel syndrome. *Neurogastroenterol Motil*. 2015;27(8):1110–1119.
29. Simsek I, Meric C, Erdem H, Pay S, Kilic S, Dinc A. Accuracy of recall of the items included in disease activity forms of Behcet's disease: comparison of retrospective questionnaires with a daily telephone interview. *Clin Rheumatol*. 2008;27(10):1255–1260.
30. Stewart WF, Lipton RB, Simon D, Liberman J, Von Koff M. Validity of an illness severity measure for headache in a population sample of migraine sufferers. *Pain*. 1999;79(2-3):291–301.
31. Stone AA, Broderick JE, Shiffman SS, Schwartz JE. Understanding recall of weekly pain from a momentary assessment perspective: absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain*. 2004;107(1-2):61–69.
32. Stone AA, Broderick JE, Schwartz JE. Validity of average, minimum, and maximum end-of-day recall assessments of pain and fatigue. *Contemp Clin Trials*. 2010;31(5):483–490.
33. Stone AA, Broderick JE, Junghaenel DU, Schneider S, Schwartz JE. PROMIS fatigue, pain intensity, pain interference, pain behavior, physical function, depression, anxiety, and anger scales demonstrate ecological validity. *J Clin Epidemiol*. 2016;74:194–206.
34. Tran BR, Thomas AG, Vaida F, et al. Comparisons of reported sexual behaviors from a retrospective survey versus a prospective diary in the Botswana Defence Force. *AIDS Educ Prev*. 2013;25(6):495–507.
35. Wood WA, Deal AM, Bennett AV, et al. Comparison of seven-day and repeated 24-hour recall of symptoms in the first 100 days after hematopoietic cell transplantation. *J Pain Symptom Manag*. 2015;49(3):513–520.
36. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *J R Stat Soc*. 1983;32(3):307–317.
37. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
38. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30–46.
39. Weinfurt KP. Constructing arguments for the interpretation and use of patient-reported outcome measures in research: an application of modern validity theory. *Qual Life Res*. 2021;30(6):1715–1722.