Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Comparison of 7-Day Recall and Daily Diary Reports of COPD Symptoms and Impacts

Antonia V. Bennett, PhD[1,*], *Dagmar Amtmann, PhD*[2], *Paula Diehr, PhD*[3,4], *Donald L. Patrick, PhD, MSPH*[3]

[1]*Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA;* [2]*Departments of Rehabilitation Medicine,* [3]*Health Services, and* [4]*Biostatistics, University of Washington, Seattle, WA, USA*

ABSTRACT

Objective: Patient reporting of symptoms in a questionnaire with a 7-day recall period was expected to differ from symptom reporting in a 7-day symptom diary on the basis of cognitive theory of memory processes and several studies of symptoms and health behaviors. **Methods:** A total of 101 adults with chronic obstructive pulmonary disease (COPD) completed a daily diary of items measuring symptoms and impacts of COPD for 7 days, and on the seventh day they completed a questionnaire of the same items with a 7-day recall period. The analysis examined concordance of 7-day recall with summary descriptors of the daily responses, examined the magnitude and covariates (patient characteristics and response patterns) of the difference between 7-day recall and mean of daily responses, and compared the discriminant ability and ability to detect change of 7-day recall and mean of daily responses. **Results:** A 7-day recall was moderately concordant with the mean and maximum of daily responses and was 0.34 to 0.50 SDs higher than the mean of daily responses. Only the weekly report itself was a covariate of the difference. The discriminant ability and ability to detect change were equivalent. **Conclusions:** In measuring the weeklong experience of COPD symptoms and impacts on groups of patients, the 7-day recall scores were higher than the daily diary scores, but equivalent in detecting change over time.

Keywords: COPD, mental recall, questionnaires, signs and symptoms, validation studies.

Introduction

Dyspnea is one of the most common and disabling symptoms for patients with chronic obstructive pulmonary disease (COPD) [1]. The American Thoracic Society's consensus statement on dyspnea describes dyspnea as "a subjective experience of breathing discomfort that consists of qualitatively distinctive sensations that vary in intensity" [2]. Patients with COPD frequently report reductions in daily functional status and overall quality of life because of limitations in their activities due to respiratory discomfort. Assessment of the patient experience is especially important in COPD, because there is known discrepancy between the severity of COPD and dyspnea per clinical assessment and the severity of dyspnea and its impacts experienced by patients [2,3].

The Food and Drug Administration Patient-Reported Outcome (PRO) Guidance describes the standardized process of developing PRO instruments for clinical trials. One step is selecting and providing justification for the recall period [4]. A model of recall periods in PRO instruments illustrates that the optimal recall period depends in part on the characteristics of the event being measured, such as the frequency of occurrence and the rate of fluctuation [5]. The goal of this analysis was to compare daily diary and weekly reporting to determine whether they provide similar information about the weeklong symptom experience of patients with COPD.

The accuracy of short-term retrospective recall (1–4 weeks) in measuring daily experience has been tested in several areas: urinary and fecal incontinence [6–9], symptoms and impacts of type 2 diabetes [10], symptoms of cystic fibrosis [11], physical activity [12,13], and alcohol consumption [14,15]. Correlation between daily diary and retrospective reporting varied widely across studies, ranging from 0.33 to 0.89. Furthermore, within some studies the correlation varied by level of bother or severity and by patient characteristics. A study of urinary incontinence found that the correlation was higher for patients who were moderately or greatly bothered by urge incontinence than for those who were slightly or not bothered by it (0.812 vs. 0.528) [7]. In a study of physical activity, the correlation was higher for very hard physical activity than for less intense physical activity, and correlations were higher for men than for women [12]. Most studies of recall periods report only correlation coefficients, but not the slope of the relationship between each type of measurement or the difference in mean scores. A study of cystic fibrosis symptoms found that the concordance of 7-day recall and mean of daily diary responses was high (range: 0.72–0.85; concordance is described in the "Methods" section) and the average difference was less than one-quarter of a response scale point [11]. A study of symptoms and impacts of type 2 diabetes found moderate to high concordance of 7-day recall and mean of daily diary responses (range: 0.65–0.86) and the average difference ranged from 0.22 to 0.77 on an 11-point scale;

[*] *Address correspondence to*: Antonia V. Bennett, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, 2nd Floor, New York, NY 10065, USA.

E-mail: bennetta@mskcc.org.

1098-3015/$36.00 – see front matter Copyright © 2012, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

doi:10.1016/j.jval.2011.12.005

Table 1 – Daily and weekly items from the DPD, listed by construct.

Severity of breathing problems	
Daily item	Rate the overall severity of your breathing problems since you completed the diary this morning.
Weekly item	Rate the overall severity of your breathing problems during the past week.
Response options for both items	Not at all severe/a little severe/somewhat severe/moderately severe/severe/very severe/extremely severe
Activity limitation due to breathing problems	
Daily item	How much have your breathing problems limited you in doing the things you wanted to do today?
Weekly item	Overall, how much have your breathing problems limited you in doing the things you wanted to do during the past week?
Response options for both items	Not at all/a little/somewhat/moderately/a lot/very much/extremely
Feeling upset during breathing problems	
Daily item	How often did you feel upset when you had breathing problems today?
Weekly item	How often did you feel upset when you had breathing problems during the past week?
Response options for both items	None of the time/hardly any of the time/a little of the time/some of the time/a good bit of the time/most of the time/all of the time

Note: Each set of response options is scored from 0 to 6, where not at all or none = 0 and extremely or all of the time = 6.
DPD, Dyspnea Patient Diary.

for some items, the amount of difference was positively associated with the variation in daily reports [10].

Retrospective recall has also been compared with *ecological momentary reporting*, that is, repeated real-time reporting by subjects during their day-to-day life. These studies found that 1) recalled pain is most similar to the maximum and last reports, 2) patients with greater variability in real-time reports of pain will recall a higher level of pain than the average of their real-time reports, and 3) while the group-level correspondence between real-time reports and recall is moderate, the within-person correspondence is low [16–18].

This study is the first to compare recall periods for measuring symptoms of COPD. It is not known whether the similarities in retrospective and daily reporting found previously in respiratory symptoms (e.g., cystic fibrosis [11]) are consistent across respiratory symptoms of other diseases. This study will also provide additional data on the variation of the difference in 7-day recall and the mean of daily diary responses by patient characteristics, symptom severity, and daily variation in symptoms.

The analysis was organized in four parts: 1) comparison of the weekly item response with summary descriptors of daily diary item responses; 2) examination of the difference between the weekly response and mean of daily diary responses for subgroups defined by patient characteristics and by patterns of diary responses; 3) comparison of the ability of the weekly response and the mean of daily responses to discriminate between known groups defined by patient health status; and 4) comparison of the ability of the weekly response and the mean of daily responses to detect change over time. Two hypotheses based on cognitive theory and studies of recall were tested. In Part 1 of the analysis, it was expected that the weekly response would be most similar to the maximum daily report and the last daily response (day 7). In Part 2 of the analysis, it was expected that the difference between the weekly response and mean of daily responses would be larger in diaries with greater variability in daily responses.

Methods

This study is a secondary analysis of PRO data originally collected for the development and validation of a COPD symptom diary, which has provided a rare opportunity to compare daily diary reporting with 7-day recall.

Diary

The Dyspnea Patient Diary (DPD) assesses “relief from dyspnea” and the degree to which this relief impacts patients with COPD. It includes symptom and impact questions relevant to COPD that measure the pattern, severity, and intensity of symptoms and symptom-related impacts. The DPD is presented as an electronic handheld device (similar to a PalmPilot) with one question per screen. It is completed by patients twice each day, once in the morning just after their morning routine and before the “start” of their day, and once in the evening before going to bed. Scoring of a daily item in the DPD requires the daily item to be completed by the patient at least 5 out of 7 days.

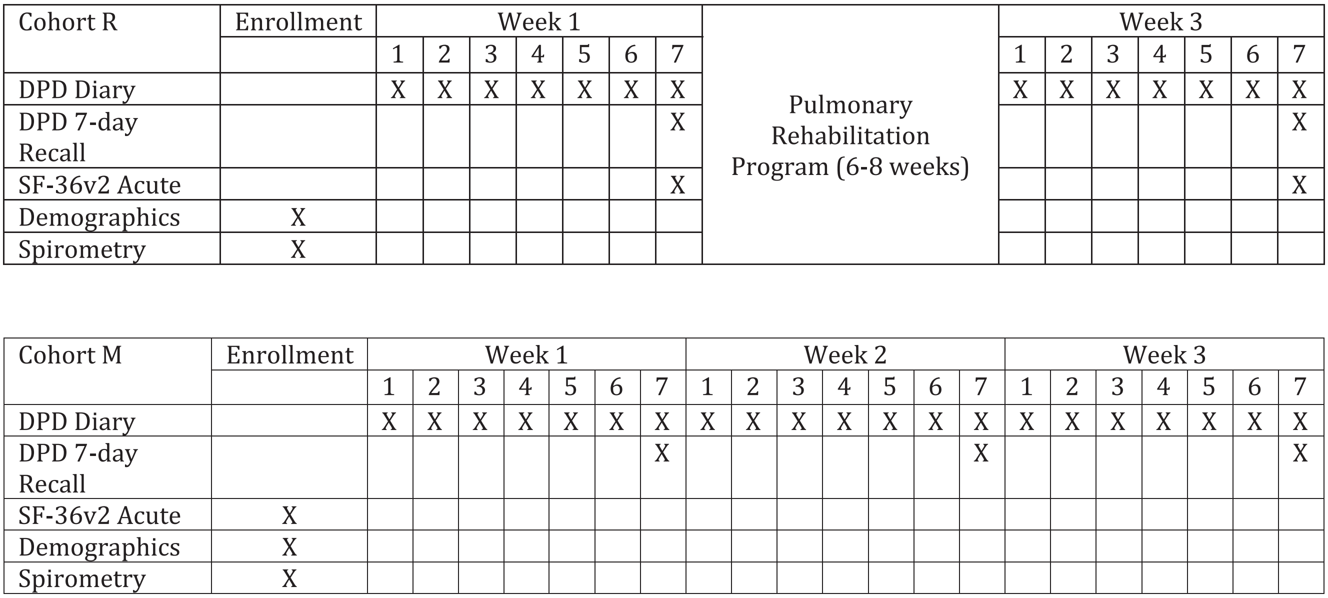
The DPD was developed by integrating information from patients, the literature, and key clinical and methodological experts. The DPD was found to have sound cross-sectional validity (unpublished report) [19]. Change in DPD scores due to treatment remains to be evaluated in a clinical trial with a sample size larger than that in the preliminary validation study.

The DPD items included in this analysis are listed in Table 1; the three constructs measured by the items are *overall severity of breathing problems*, *activity limitations due to breathing problems*, and *feeling upset during breathing problems*. These three constructs were measured by the DPD by using a daily item administered in the evening that asks about that day’s experience and a weekly item with a 7-day recall period that is completed on the last day (day 7) of the diary period.

The DPD assessed the daytime but not the nighttime severity of these three constructs in the daily diary. Data on the occurrence of distinct breathing problems (e.g., inability to breathe; difficult, heavy breathing; and trouble breathing deeply) measured in the morning and evening diaries indicate an equal or slightly higher, but no statistically significant difference in the rate of breathing problems in the daytime compared with the nighttime [19]. We expect that the daytime and nighttime severities of symptoms are the same to the point that the daytime report is representative of the daytime and nighttime average.

Data collection

Data were collected in an observational study of 101 patients with COPD beginning a new treatment as part of a natural progression in their medical care. Research participants were recruited through medical clinics in the United States at the point when their treating clinician believed that their current stage of COPD



Cohort R	Enrollment	Week 1							Pulmonary Rehabilitation Program (6-8 weeks)	Week 3						
		1	2	3	4	5	6	7		1	2	3	4	5	6	7
DPD Diary		X	X	X	X	X	X	X		X	X	X	X	X	X	X
DPD 7-day Recall								X								X
SF-36v2 Acute								X								X
Demographics	X															
Spirometry	X															

Cohort M	Enrollment	Week 1							Week 2							Week 3						
		1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
DPD Diary		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
DPD 7-day Recall								X							X							X
SF-36v2 Acute	X																					
Demographics	X																					
Spirometry	X																					

Fig. 1 – Schedule of assessments.

necessitated a change in therapy. Eligible patients were male or female, ages ranging from 40 to 80 years, diagnosed with COPD characterized by airflow limitation that is not fully reversible, met the spirometry criteria for moderate to very severe stage of disease (forced expiratory volume in the first second of expiration [FEV_1]/ forced vital capacity <70% and FEV_1 <80% predicted) [20], current or ex-smokers with a smoking history of 10 or more pack-years, and had self-reported breathing difficulty (Medical Research Council [MRC] grade ≥1; the MRC scale is described below). Patients were excluded if they used oxygen therapy at rest as well as during daily activities (this did not include those who used it solely during sleep or those who used it for exertion); if they had experienced three or more COPD exacerbations that required inpatient hospitalization in the past 12 months; if they had active cardiovascular comorbidity within the past 6 months that was unstable and/or contributed significantly to their dyspnea; if they had self-reported or a physician diagnosis of concurrent asthma or respiratory allergies; if they had a history of lung cancer, lung resection, lung volume reduction surgery, or pulmonary fibrosis; and if they had an admitted significant history of alcohol or drug abuse within the past year. The partial list of inclusion and exclusion criteria listed above is intended to describe the health status of the study sample in this analysis; a complete list is available on request.

Patients were then assigned to either cohort R, a pulmonary rehabilitation program lasting 6 to 8 weeks, or cohort M, in which patients received a pharmacological maintenance medication for COPD, depending on the patient's needs.

Figure 1 shows the measurement schedule for each cohort. Cohort R patients completed the DPD for 7 days before the start of their pulmonary rehabilitation program, and then they resumed use of the DPD for 7 additional days beginning the day after the end of the rehabilitation program. On the last day of each 7-day diary period, cohort R patients completed the DPD items about their symptom experience over the past 7 days. This provided a maximum of 2 weeks of diary data for analysis for patients in cohort R. Cohort M patients completed the DPD every day over the course of 4 weeks. The change in their medication began on the second day of diary completion. The majority of cohort M patients returned the DPD PalmPilot at their follow-up clinic visit that occurred, as possible, near the end of week 4. Therefore, only a maximum of 3 weeks of diary data were analyzed for cohort M patients.

The short-form 36 health survey version 2 (SF-36v2) Acute version was completed by cohort R at the end of weeks 1 and 2 and by cohort M at baseline. The SF-36 Acute version, US version 2.0, is a 36-item measure of general health status. The acute version of the SF-36 has a recall period of the past 7 days instead of the past 4 weeks. The Physical Functioning and Mental Health subscale scores were included in this analysis, and are scored on a scale of 0 to 100, in which the US average is 50 and higher scores indicate better health [21,22].

Patient demographic information, including age, gender, ethnicity, and educational level, and disease-specific health information, including years since diagnosis, smoking history, lung function, and MRC grade, were collected at enrollment. Lung function was measured via spirometry test and reported as FEV_1%. FEV_1% indicates forced expiratory volume in the first second as a percent of the volume predicted for a person with healthy lung function of the same age, height, and sex. Patients with FEV_1 between 30% and 50% predicted are limited in daily activity by shortness of breath, and patients with FEV_1 of less than 30% have severe functional impairment [23]. The MRC scale is a single-item measure in which patients report the amount of activity that produces dyspnea; the response options are 0 = "strenuous exertion," 1 = "hurrying on level ground," 2 = "stop after a mile or so," 3 = "stop after 100 yards," and 4 = "can't leave the house" [2,24].

Human subjects research

Institutional review board approval was granted for the collection of data and additionally for the secondary use of de-identified data in this analysis; data collection and analysis were conducted in accordance with ethical standards described in the 1964 Declaration of Helsinki.

Statistical analyses

Analyses were limited to observations with both the weekly response and at least five nonmissing daily responses. Each of

Table 2 – Descriptors of the daily diary item responses.

Descriptor	What it measures
Mean	Average of daily reports
Median	Median of daily reports
Mode—max	Most common response (largest mode if multiple modes)
Mode—min	Most common response (smallest mode if multiple modes)
Maximum	Most severe response
Minimum	Least severe response
Day 7 (last day)	Response of the same day the weekly report was completed
Day 6	Response of the day before the weekly report was completed
Day 1 (first day)	First daily report

Note: The weekly report comprises items with 7-day recall completed on day 7 of the weeklong diary period.

the three constructs was analyzed separately, because constructs that are emotionally charged, such as *feeling upset during breathing problems*, may be recalled differently. Statistical significance was defined as $P < 0.05$ with two-sided 95% confidence intervals, and standard errors were estimated by using the clustered sandwich estimator to account for correlation from multiple diaries per patient.

1. *Comparison of the weekly response with descriptors of the daily diary responses*

For each construct, the weekly response was compared with descriptors of the daily diary responses (e.g., mean, median, and maximum of daily responses) to identify what the weekly response most closely approximated. A list of each descriptor is shown in Table 2. The weekly response was compared with each descriptor of the daily diary response by the difference in group means and by a measure of concordance. The difference in group means was evaluated on the basis of whether it was less than a small effect size of 0.20 SDs [25]. This effect size was calculated by dividing the difference by the average of the two SDs (SD mean daily diary and SD weekly response), because there was no reason to choose one over the other. Concordance was measured by the concordance correlation coefficient (CCC), which is composed of Pearson's correlation coefficient and a bias correction factor. The bias correction factor ranges from 0 to 1, where values closer to 1 indicate less deviation of the Pearson's best-fit line from the 45° origin line (the concordance line) [26,27]. The CCC ranges from −1 to +1, and values closer to 1 indicate greater concordance. The CCC was interpreted per the threshold of agreement recommended for the intraclass correlation coefficient, in which 0.70 or greater indicates a high level of agreement, because the statistics are algebraically similar and consistent with each other [26]. The intraclass correlation coefficient has the disadvantage that it cannot be decomposed into a correlation coefficient and bias factor.

2. *Examine the difference between the weekly response and the mean of daily diary responses for subgroups defined by patient characteristics and patterns of diary responses*

The amount of difference between the weekly response and mean daily responses was compared by subgroups of patient characteristics and response patterns to identify characteristics of the data in which the results of daily diary and weekly measurement may differ. The patient characteristics and subgroups were gender, age (49–64 years, ≥65 years), education (beyond high school and not beyond high school), tertiles of SF-36v2 Acute Physical Functioning and Mental Health subscale scores, MRC grade (grades 0–1, 2, and 3–4), and years since diagnosis with COPD (<2 years and >2 years). The response patterns were the range of daily responses (maximum minus minimum daily response), the SD of daily responses, the mean of daily responses, and the weekly response. For each characteristic, to evaluate variation in the amount of difference by subgroup (or by units of continuous variables), linear regression was used to test for a trend where the null hypothesis was that the slope of the trend line is zero. The analyses in Parts 2, 3, and 4 of this study compare weekly response with the mean of daily responses because diaries measuring symptom severity are typically scored on the basis of the mean of daily reports.

3. *Comparison of the ability of the weekly response and the mean of daily diary responses to detect differences between known groups*

The ability of the mean of daily diary responses and the weekly response to detect differences between known groups (i.e., discriminant ability) was compared by calculating *t* statistics via linear regression for the differences in item scores by sample subgroups and comparing the *t* statistics between each type of measurement. The subgroups were defined by tertiles of SF-36 Physical Functioning subscale scores, tertiles of SF-36 Mental Health subscale scores, and MRC grade (0–1, 2, and 3–4). The null hypothesis of the *t* statistic was that there was no difference between sample subgroups. The discriminant ability was not tested by using subgroups of FEV_1% because spirometry values are known to be weak predictors of functional status and quality of life [28].

4. *Comparison of the ability of the weekly response and the mean of daily diary responses to detect change over time*

The average change in scores from time 1 to time 2 was compared between the weekly questionnaire and the daily diary. Diaries at time 1 were the first diaries (week 1) of both cohort R and cohort M. Diaries at time 2 were the last diaries for each cohort (week 2 for cohort R and week 3 for cohort M). For each type of measurement, a *t* statistic was calculated for the difference in mean scores between time 1 and time 2, where the null hypothesis was that the difference is zero. The effect size of the change from time 1 to time 2, calculated as Cohen's *d* and using pooled SD, was also determined [25,29].

Results

Patient characteristics

Eight clinic sites located in the United States recruited and enrolled a total of 101 patients in the study. Diary data from 98 (97%) patients were used in this analysis; 3 patients did not have sufficient daily or weekly data to have at least 1 week of diary data included in the analysis. Of the 98 patients, 73 (74%) were enrolled through sites in California, 17 (17%) through sites in Pennsylvania and Michigan, and 8 (8%) through sites in Georgia and Alabama. Patient demographic characteristics are shown in Table 3. Fifty-two (53%) patients were females, and 94 (96%) patients were white. The mean (SD) age of patients was 65.2 (7.7) years, and the mean (SD) years of education was 13.3 (2.4). Forty-eight (49%) patients were educated beyond high school. Thirty-nine (40%) patients were enrolled in cohort R, and 59 (60%) patients were enrolled in cohort M. Demographic characteristics did not differ by cohort.

Patient clinical and self-reported health

The mean (SD) time since diagnosis with COPD was 4.8 (5.6) years (Table 3). Patients had a smoking history measured by mean (SD) pack-years of 53.8 (29.3). Cohort R had more limited lung function

Table 3 – Patients' demographic characteristics.

	All patients, n (%)
	98 (100)
Females	52 (53)
Age (y), mean (SD)[*]	65.2 (7.7)
49–64	50 (51)
≥65	48 (49)
White	94 (96)
Nonwhite[†]	4 (4)
Years of education, mean (SD)[‡]	13.3 (2.4)
Beyond HS	48 (49)
Not beyond HS	50 (51)
Years since COPD diagnosis	4.8 (5.6)
Number of pack-years[§]	53.8 (29.3)
FEV_1 (%)	46.4 (15.6)[‖]
MRC score	2.3 (1.0)[‖]
0 = "Strenuous exertion"	0%
1 = "Hurrying on level ground"	25%
2 = "Stop after a mile or so"	36%
3 = "Stop after 100 yards"	26%
4 = "Can't leave house"	13%
Self-rated health	3.4 (0.9)
1 = Excellent	1%
2 = Very Good	15%
3 = Good	39%
4 = Fair	32%
5 = Poor	13%
SF-36v2 Acute[¶]	
Physical functioning	30.8 (9.4)[‖]
Role-physical	34.0 (10.6)
Bodily pain	44.4 (11.0)
General health	38.4 (10.5)
Vitality	41.5 (10.3)
Social functioning	40.4 (12.7)
Role-emotional	40.4 (13.7)
Mental health	45.7 (11.5)
Physical component score	34.2 (9.6)
Mental component score	46.3 (12.7)

COPD, chronic obstructive pulmonary disease; FEV_1, forced expiratory volume in the first second of expiration; HS, high school; MRC, Medical Research Council; SF-36v2, Short-Form 36 Health Survey, version 2.

[*] Patients were eligible if between 40 and 80 y old.

[†] Four patients (4%) were nonwhite; one was black/African American, one was Asian/Pacific Islander, and two were Hispanic/Latino.

[‡] Years of education is the highest grade of school completed (e.g., 12 = high school graduate, 16 = graduate of 4-y college).

[§] One pack-year is equal to smoking one pack (20 cigarettes) per day for 1 y. One patient in the medicine cohort did not have this information—the *t* test is based on *df* = 95.

[‖] Difference between cohort R and cohort M is significant at $P < 0.05$; values are reported in text.

[¶] SF-36v2 Acute version data reported are from the first assessment of each cohort.

(%FEV_1) than did cohort M (41.5 [15.5] vs. 49.7 [15.0]; $P < 0.01$). The limitations due to breathing problems, measured by the MRC scale, were greater for cohort R than for cohort M (2.9 [0.9] vs. 1.9 [0.9]; $P < 0.001$). Patient's self-rated health was low; 16% rated their health as "excellent" or "very good," 39% as "good," and 45% as "fair" or "poor." Patient's physical and mental health, as measured by the SF-36v2 Acute version, was lower than the US average; patients had a mean (SD) Physical Functioning subscale score of 30.8 (9.4) and a mean (SD) Mental Health subscale score of 45.7 (11.5). The SF-36v2 Acute Physical Functioning subscale scores were lower for cohort R than for cohort M (32.3 [23.2] vs. 41.4 [21.1]; $P < 0.049$).

Diary characteristics

Patients completed, at least partially, 252 weeks of dairy data. Eighteen (7%) of these weekly diaries were missing the weekly report and were therefore excluded from analysis. An additional eight (3%) diaries had fewer than five daily reports and were also excluded from analysis. Of the remaining 226 diaries, 161 (71%) had five daily reports and 65 (29%) had five or six daily reports. The missing days were equally distributed across the first 6 days of the diary; no diaries with the weekly report were missing the day 7 daily response. The missing days were the same across the three constructs because the electronic diary required completion of each item. The difference between the weekly response and the mean of daily responses was the same for diaries with no missing days and diaries with one or two missing days for each construct except *feeling upset during breathing problems*. For this construct, the difference was larger for diaries with no missing days by 0.29 ($P < 0.014$).

Analysis was based on 226 weeks of diary data from 98 patients, comprising 71 (31%) diaries from cohort R and 155 (69%) diaries from cohort M (Table 4). The majority of patients completed the maximum number of diaries for their cohort; two diaries were completed by 32 (82%) rehabilitation patients, and three diaries were completed by 43 (73%) medicine patients. The amount of within-person variation for each week of diary data was calculated as the SD of daily responses for an item within one diary. The average within-person variation for the item *overall severity of breathing problems* was 0.58 scale points, for *activity limitation due to breathing problems* it was 0.65 scale points, and for *feeling upset during breathing problems* it was 0.58 scale points. Table 4 lists the mean and standard error for the weekly response and the mean of daily responses for each construct.

Table 4 – Diary characteristics.

No. of diaries	226
No. of patients	98
No. of diaries per patient, mean (SD)	2.3 (0.7)
One diary	13 (13%)
Two diaries	42 (43%)
Three diaries	43 (44%)
Severity of breathing problems, mean (SE)	
Weekly severity	1.71 (0.13)
Mean of 7 days	1.13 (0.09)
Difference	0.59 (0.08)[*]
Activity limitation due to breathing problems, mean (SE)	
Weekly severity	2.02 (0.13)
Mean of 7 days	1.53 (0.13)
Difference	0.48 (0.06)[*]
Feeling upset during breathing problems, mean (SE)	
Weekly severity	1.58 (0.13)
Mean of 7 days	1.01 (0.11)
Difference	0.56 (0.06)[*]

SE, standard error.

[*] Difference between weekly severity and mean of 7 days is not zero ($P < 0.001$).

Table 5 – Comparison of weekly response with descriptors of daily response (N = 226 diaries).

	Sample		Difference		Concordance correlation coefficient		
	Mean	SD	Mean	SD	CCC ($r \times b$)	Pearson (r)	Bias factor (b)
Severity of breathing problems							
Weekly response	1.71	1.38					
Mean	1.13	0.97	0.59	0.89	0.64	0.77	0.84
Median	1.04	1.03	0.67	0.96	0.60	0.72	0.83
Mode—max	1.11	1.14	0.61	0.99	0.62	0.71	0.88
Mode—min	0.95	1.03	0.76	1.00	0.56	0.70	0.80
Maximum	1.95	1.34	−0.23	1.00	0.72	0.73	0.98
Minimum	0.50	0.80	1.21	1.09	0.34	0.62	0.55
Day 7 (last day)	1.08	1.17	0.63	1.08	0.57	0.65	0.88
Day 6	1.09	1.13	0.61	1.04	0.59	0.67	0.88
Day 1 (first day)	1.27	1.25	0.44	1.16	0.58	0.62	0.94
Activity limitation due to breathing problems							
Weekly response	2.02	1.43					
Mean	1.53	1.37	0.48	0.83	0.78	0.83	0.94
Median	1.48	1.48	0.54	0.93	0.75	0.80	0.94
Mode—max	1.56	1.59	0.46	0.99	0.75	0.79	0.95
Mode—min	1.41	1.53	0.61	0.99	0.72	0.78	0.92
Maximum	2.42	1.57	−0.40	1.00	0.75	0.78	0.96
Minimum	0.81	1.15	1.21	0.97	0.50	0.74	0.68
Day 7 (last day)	1.45	1.54	0.57	1.02	0.71	0.77	0.93
Day 6	1.52	1.54	0.49	1.13	0.67	0.71	0.95
Day 1 (first day)	1.61	1.57	0.42	1.07	0.72	0.75	0.96
Feeling upset during breathing problems							
Weekly response	1.58	1.35					
Mean	1.02	1.11	0.56	0.82	0.71	0.80	0.89
Median	0.94	1.19	0.63	0.85	0.69	0.78	0.88
Mode—max	0.98	1.26	0.60	0.93	0.68	0.75	0.90
Mode—min	0.83	1.17	0.75	0.97	0.60	0.71	0.84
Maximum	1.87	1.59	−0.29	1.01	0.75	0.78	0.97
Minimum	0.42	0.83	1.16	1.08	0.35	0.60	0.58
Day 7 (last day)	0.95	1.26	0.63	0.99	0.64	0.71	0.89
Day 6	0.95	1.25	0.64	0.98	0.64	0.71	0.89
Day 1 (first day)	1.17	1.42	0.44	1.22	0.58	0.61	0.96

CCC, concordance correlation coefficient.

Analysis of weekly and daily diary reporting

1. Comparison of the weekly response with descriptors of the daily diary responses

For each construct, the weekly response was less than the maximum daily response (the difference ranged from −0.23 to −0.40 and was nonzero; $P < 0.001$) and the weekly response was higher than the mean of daily responses (the difference ranged from 0.48 to 0.59 and was nonzero; $P < 0.01$) (Table 5). The mean (standard error) of the difference between the weekly report and the mean of daily responses was 0.59 (0.08) for *overall severity of breathing problems*, 0.48 (0.08) for *activity limitation due to breathing problems*, and 0.56 (0.06) for *feeling upset during breathing problems*; the effect sizes of the difference for each construct were 0.50, 0.34, and 0.46, respectively.

For each construct, the Pearson's correlation coefficient was larger for the mean of daily responses than it was for the maximum; however, the bias correction factor of the maximum was closer to 1.00 than it was for the mean of daily responses; therefore, the CCC, which is the product of Pearson's correlation coefficient and the bias correction factor, was larger for the maximum than it was for the mean of daily responses for two of the three constructs. For the three constructs, the CCC of the mean daily response ranged from 0.64 to 0.78 and the CCC of the maximum daily response ranged from 0.72 to 0.75.

2. Examine the difference between the weekly response and the mean of daily diary responses for subgroups defined by patient characteristics and patterns of diary responses

The difference between the weekly response and the mean of daily responses for each construct was compared by subgroups of patient demographics and health status. The difference was greater for females than for males for the construct *activity limitation due to breathing problems* (0.60 vs. 0.35; $P < 0.026$), and the difference was greater for younger patients (age 49–64 years vs. ≥65 years) for the construct *feeling upset during breathing problems* (0.70 vs. 0.42; $P < 0.03$). The difference was larger with lower SF-36v2 Acute Physical Functioning scores for *overall severity of breathing problems* (low: 1.09, middle: 0.43, high: 0.36; $P < 0.003$). There was no variation in the difference by education level, SF-36v2 Acute Mental Health score, MRC grade, or years since diagnosis.

The difference between weekly and mean daily responses for each construct was also compared by the range, SD, and mean of daily responses, as well as by the weekly response (Table 6). The difference increased with a larger range ($P < 0.015$) and a larger SD ($P < 0.012$) in daily responses only for the construct *overall severity of breathing problems*. The difference decreased by the mean of daily responses for *activity limitation due to breathing problems* ($P < 0.015$). For each construct, the difference increased with higher weekly responses ($P < 0.001$).

Table 6 – Difference between weekly and mean daily reporting for each construct, by patient response.

Difference	Severity of breathing problems		Activity limitation due to breathing problems		Feeling upset during breathing problems	
	Diaries	Mean (SE)	Diaries	Mean (SE)	Diaries	Mean (SE)
Total	226	0.59 (0.08)	226	0.48 (0.06)	226	0.56 (0.06)
Range of daily responses						
0	39	0.26 (0.08)	22	0.23 (0.13)	59	0.32 (0.08)
1	91	0.62 (0.10)	98	0.55 (0.07)	68	0.50 (0.10)
2	62	0.48 (0.11)	66	0.58 (0.11)	61	0.75 (0.12)
3	26	1.17 (0.28)	28	0.29 (0.24)	22	0.73 (0.20)
4	6	0.71 (0.39)	10	0.19 (0.28)	9	1.22 (0.27)
5	2	0.43 (0.20)	2	1.21 (1.17)	6	0.74 (0.23)
6	0	–	0	–	1	−3.43 (0.00)
t (sig.)*		2.47 (0.015)		−0.22 (0.827)		1.54 (0.126)
SD of daily responses						
Per unit change†	226	0.47 (0.18)	226	−0.03 (0.14)	226	0.25 (0.15)
t (sig.)*		2.56 (0.012)		−0.23 (0.817)		1.65 (0.103)
Mean of daily responses						
Per unit change†	226	0.09 (0.07)	226	−0.13 (0.05)	226	−0.02 (0.06)
t (sig.)*		1.28 (0.205)		−2.48 (0.015)		−0.45 (0.651)
Weekly response						
0 (Not at all)	43	−0.16 (0.04)	29	−0.37 (0.17)	62	−0.15 (0.04)
1	76	0.27 (0.05)	66	0.37 (0.05)	58	0.44 (0.10)
2	47	0.80 (0.09)	59	0.79 (0.10)	44	0.90 (0.09)
3	36	0.82 (0.17)	34	0.53 (0.16)	41	1.16 (0.17)
4	13	1.65 (0.22)	22	0.73 (0.22)	17	1.19 (0.21)
5	9	2.78 (0.55)	14	0.93 (0.26)	4	0.68 (0.39)
6 (Extremely)	2	2.50 (0.56)	2	1.64 (0.56)	0	–
t (sig.)*		7.59 (<0.001)		3.93 (<0.001)		7.89 (<0.001)

Note: Each cell reports the mean (SE) of the difference between weekly and mean daily report.

SE, standard error.

* The *t* statistic is reported from the *t* test of regression coefficient indicating the variation in the difference by this variable ($df = 97$); *P* value for the alternate hypothesis of nonzero difference is shown.

† *Per unit change* is the change in difference per unit change in the mean of daily responses.

3. *Comparison of the ability of the weekly response and the mean of daily diary responses to detect differences between known groups*

The discriminant ability of the weekly response and the mean of daily diary responses was compared for subgroups of the SF-36v2 Acute Physical Functioning and Mental Health subscales and MRC grade. Therefore, discriminant ability was compared in nine instances (three constructs × three patient variables). In each comparison, the standard error was consistently smaller for the mean of daily diary responses than for the weekly response. The absolute value of the *t* statistic for a nonzero linear trend across subgroups was larger for the mean of daily diary responses than for the weekly response in seven out of the nine comparisons. The absolute value of the *t* statistic was smaller for the mean of daily responses than for the weekly response for *overall severity of breathing problems* by SF-36v2 Acute Physical Functioning subgroups and for *feeling upset during breathing problems* by MRC grade. Nevertheless, in every comparison, the significance of the *t* statistic was the same.

4. *Comparison of the ability of the weekly response and the mean of daily diary responses to detect change over time*

The average change from time 1 to time 2 was estimated for the weekly response and for the mean of daily responses (Table 7). The *t* statistic for a nonzero change between time 1 and time 2 is presented. For each construct, the absolute value of the *t* statistic was larger for the mean of daily responses than it was for the weekly response; however, the significance of the change was the same. In addition, there was very little difference in the effect sizes of the weekly response and the mean of daily responses.

Discussion

For each construct, the weekly response was consistently lower than the maximum daily response by about one-quarter a response scale point ($P < 0.01$) and it was higher than the mean of the daily responses for that diary period by about one-half a response scale point ($P < 0.001$). The difference between the mean of daily responses and the weekly response ranged from 0.34 to 0.50 SDs, which is too large to be considered equivalent. The weekly response was most concordant with the mean and the maximum of daily responses. The hypothesis that the weekly response would be most similar to the maximum and the last daily response was not confirmed. The concordance of the weekly response and the mean of daily responses was moderate, ranging from 0.64 to 0.78.

There was very little variation in difference between weekly response and the mean of daily responses by patient characteristics. The difference was greater for females for the construct *activity limitation due to breathing problems* ($P < 0.026$), it was greater for younger patients (age 49–64 years vs. ≥65 years) for the construct *feeling upset due to breathing problems* ($P < 0.03$), and it was greater with lower SF-36v2 Acute Physical Functioning score for the construct *overall severity of breathing problems* ($P < 0.003$). There was no variation in the difference for any construct by education level,

Table 7 – Difference by change over time.

	Mean (SE)			Effect size*	t (sig.)†
Construct	Time 1 (diaries = 91)	Time 2 (diaries = 87)	Change		
Severity of breathing problems					
Weekly report	2.01 (0.15)	1.56 (0.14)	−0.45 (0.12)	0.32	−3.80 (<0.001)
Mean of 7 days	1.38 (0.11)	1.02 (0.10)	−0.35 (0.08)	0.36	**−4.45** (<0.001)
Difference	0.63 (0.10)	0.54 (0.09)			
Activity limitation due to breathing problems					
Weekly report	2.24 (0.16)	1.89 (0.15)	−0.36 (0.13)	0.25	−2.67 (0.009)
Mean of 7 days	1.82 (0.16)	1.43 (0.14)	−0.39 (0.12)	0.28	**−3.29** (0.001)
Difference	0.42 (0.10)	0.46 (0.07)			
Feeling upset during breathing problems					
Weekly report	1.80 (0.15)	1.49 (0.15)	−0.31 (0.12)	0.22	−2.61 (0.011)
Mean of 7 days	1.12 (0.12)	0.96 (0.12)	−0.26 (0.09)	0.23	**−2.78** (0.007)
Difference	0.58 (0.08)	0.54 (0.10)			

Note: Time 1 is the first diary (week 1) from the rehabilitation and medicine cohorts. Time 2 is the last diary from each cohort, that is, week 2 diary of the rehabilitation cohort and week 3 diary of the medicine cohort. All diaries are included except medicine cohort week 2 diaries ($d = 48$). SE, standard error.

* The effect size is calculated as Cohen's d by using pooled SD.

† The t statistic is reported from the t test of regression coefficient indicating the variation in the patient response by this variable ($N = 97$, $df = 96$); P value for the alternate hypothesis of nonzero difference is shown. To highlight any differences in discriminant ability between daily diary and weekly reporting, the t statistic with the larger absolute value is in **bold**.

SF-36v2 Acute Mental Health score, MRC grade, or years since diagnosis.

The difference increased with a larger range ($P < 0.015$) and larger SD ($P < 0.012$) in daily responses only for the construct *overall severity due to breathing problems*. Therefore, there was not sufficient evidence to confirm the hypothesis that the difference would be larger in diaries with a larger range of daily responses. The difference decreased with larger mean of daily responses only for the construct *activity limitation due to breathing problems* ($P < 0.015$). For each construct, the difference between the weekly response and the mean of daily responses was larger for higher weekly responses ($P < 0.001$). These findings suggest that the magnitude of the difference between the weekly response and the mean of daily responses was primarily driven by the weekly response itself, and the difference was not associated with patient characteristics, symptom severity, or variability of symptoms measured in the daily diary in any consistent way.

The discriminant ability of the weekly report and the daily diary was equivalent in this study sample in that the significance of P values was consistent in each comparison. The ability to detect change over time was also consistent between the weekly report and the daily diary; the significance of P values was the same and the effect sizes were very similar. All observations at time 1 and time 2 were included in the change analysis; therefore, patients who were no longer in the study at time 2 contributed an observation only at time 1. Although loss to follow-up can bias the measurement of change, the weekly report and the mean of daily diary reports were from the same set of time 1 and time 2 observations, and so the loss to follow-up was identical.

One strength of this study was that all weekly and daily diary responses were captured via an electronic handheld device that recorded the date and time of all responses. This approach made it possible to confirm that responses were made within the correct time frame. Furthermore, the layout and wording of the questionnaire items in the weekly response were identical to the corresponding items in the daily diary except for the recall period.

This study compared daily diary and weekly reporting, pooling data across the medication and rehabilitation treatment cohorts, because there was no reason to expect that the treatment type itself would affect recall beyond the effect of patient characteristics and response patterns. The findings from the analysis of change over time in the pooled data were also evident in each cohort analyzed separately.

To test whether completing the daily diary affected the patients' weekly response, a study of weekly and daily diary responses in patients with type 2 diabetes randomized patients into two groups in which one group completed the daily diary and the weekly report and the other completed only the weekly report. It was found that the differences in the weekly responses between groups were very small and not statistically significant [10].

The relationship between these two types of symptom measurement may be different for other patient populations or in measurement of other types of symptoms and events. Smaller differences between the mean of daily diary reports and 7-day recall were found in analyses of data from youth and adults with cystic fibrosis and adult with type 2 diabetes [10,11].

We do not know the prevalence in this data of exacerbations confirmed by clinicians. An exacerbation would increase the severity of daily symptoms, and if it started midweek, it would increase the variability of symptom severity in the diary. We examined the effect of both the level of symptoms (mean of daily reports) and the variability (range of daily reports) on recall and found that these were not consistent covariates of the difference between the average of the daily diary and 7-day recall.

In choosing between a daily diary and weekly reporting, one consideration is the extra burden on patients of completing a questionnaire daily instead of once a week. If the outcome of interest were day-to-day variation or the exact timing of the onset of a respiratory exacerbation, then only the daily diary would be suitable. Nevertheless, to estimate the average experience of a group of patients during a 7-day period, the results of this study draw into question the added value of a daily diary. Although scores from 7-day recall were higher than those from the daily diary, these two types of measurements were very consistent in detecting differences between known groups and change from one diary period to another.

Acknowledgment

We thank Donald Bushnell, MA, Health Research Associates, Inc., for providing secondary data files and orientation to the data set.

Source of financial support: No funding was received for this secondary analysis or manuscript preparation. Data were collected in an observational study sponsored by Pfizer, Inc., in contract with Health Research Associates, Inc., and Donald Patrick.

REFERENCES

[1] Mulrow CD, Lucey CR, Farnett LE. Discriminating causes of dyspnea through clinical examination. J Gen Intern Med 1993;8:383–92.

[2] American Thoracic Society. Dyspnea. Mechanisms, assessment, and management: a consensus statement. Am J Respir Crit Care Med 1999; 159:321–40.

[3] Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res 2000;9:887–900.

[4] FDA. Patient reported outcome (PRO) measures: use in medical product development to support labeling claims. Fed Reg 2009;74:65132–33.

[5] Stull DE, Leidy NK, Parasuraman B, Chassany O. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Curr Med Res Opin 2009;25:929–42.

[6] Homma Y, Ando T, Yoshida M, et al. Voiding and incontinence frequencies: variability of diary data and required diary length. Neurourol Urodyn 2002;21:204–9.

[7] Kenton K, Fitzgerald MP, Brubaker L. What is a clinician to do—believe the patient or her urinary diary? J Urol 2006;176:633–5; discussion 5.

[8] Elser DM, Fantl JA, McClish DK. Comparison of "subjective" and "objective" measures of severity of urinary incontinence in women. Program for Women Research Group. Neurourol Urodyn 1995;14: 311–6.

[9] Fisher K, Bliss DZ, Savik K. Comparison of recall and daily self-report of fecal incontinence severity. J Wound Ostomy Continence Nurs 2008; 35:515–20.

[10] Bennett AV, Patrick DL, Bushnell DM, et al. Comparison of 7-day and repeated 24-hour recall of type 2 diabetes. Qual Life Res 2011; 20:769–77.

[11] Bennett AV, Patrick DL, Lymp JF, et al. Comparison of 7-day and repeated 24-hour recall of symptoms of cystic fibrosis. J Cyst Fibros 2010;9:419–24.

[12] Richardson MT, Ainsworth BE, Jacobs DR, Leon AS. Validation of the Stanford 7-day recall to assess habitual physical activity. Ann Epidemiol 2001;11:145–53.

[13] Stel VS, Smit JH, Pluijm SM, et al. Comparison of the LASA Physical Activity Questionnaire with a 7-day diary and pedometer. J Clin Epidemiol 2004;57:252–8.

[14] Gmel G, Daeppen JB. Recall bias for seven-day recall measurement of alcohol consumption among emergency department patients: implications for case-crossover designs. J Stud Alcohol Drugs 2007;68: 303–10.

[15] Hilton ME. A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. Br J Addict 1989; 84:1085–92.

[16] Stone AA, Broderick JE, Kaell AT, et al. Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritics? J Pain 2000;1:212–7.

[17] 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;107:61–9.

[18] Stone AA, Schwartz JE, Broderick JE, Shiffman SS. Variability of momentary pain predicts recall of weekly pain: a consequence of the peak (or salience) memory heuristic. Pers Soc Psychol Bull 2005;31: 1340–6.

[19] Martin ML, Bushnell D, Hareendran A, Rudell K. Report on the Validation of a Patient Reported Outcome Measure for Dyspnea in Patients with COPD. Health Research Associates, Inc, and Pfizer Ltd., 2008.

[20] Pauwels RA, Buist AS, Calverley PM, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. Am J Respir Crit Care Med 2001;163:1256–76.

[21] Ware JE Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36), I: conceptual framework and item selection. Med Care 1992;30:473–83.

[22] Ware JE Jr., Kosinski M, Dewey JE. How to Score Version 2 of the SF-36 Health Survey. Lincoln, RI: Quality Metric, Inc., 2000.

[23] National Collaborating Centre for Chronic Conditions. Chronic obstructive pulmonary disease: national clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and secondary care. Thorax 2004;59(Suppl. 1):181–272.

[24] Fletcher CM, Elmes PC, Wood CH. The significance of respiratory symptoms and the diagnosis of chronic bronchitis in a working population. BMJ 1959;1:257–66.

[25] Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Erlbaum, 1988.

[26] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991;12(4, Suppl.):142S–58S.

[27] Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68.

[28] Jones PW. Health status measurement in chronic obstructive pulmonary disease. Thorax 2001;56:880–7.

[29] Hartung J, Knapp G, Sinha B. Statistical Meta-Analysis with Applications. Hoboken, NJ: John Wiley and Sons, 2008.