Qual Life Res (2011) 20:769–777
DOI 10.1007/s11136-010-9791-5

Comparison of 7-day and repeated 24-h recall of type 2 diabetes

**A. V. Bennett · D. L. Patrick · D. M. Bushnell ·
C. F. Chiou · P. Diehr**

Accepted: 1 November 2010 / Published online: 19 November 2010

Abstract

Purpose Patient reporting of type 2 diabetes symptoms in a questionnaire with a 7-day recall period was expected to be different from symptom reports using a 7-day diary with repeated 24-h recall based on cognitive theory of memory processes and prior literature. This study compared these two types of recall in patients diagnosed with type 2 diabetes (T2D).

Methods One hundred and forty adults with T2D completed a daily diary for 7 days containing 9 T2D-related symptom and impact items. On day 7, patients completed the same items with a 7-day recall period. We examined the concordance of 7-day recall with summary descriptors of the daily reports and compared the scores and the discriminant ability of 7-day recall and mean of daily reports.

Results Seven-day recall was most concordant with the mean of daily reports. The average difference in scores was small (range 0.22–0.77 on 11-point scale) and less than 0.5 standard deviations. For some items, the difference was positively associated with the variation in daily reports. The discriminant ability was comparable.

Conclusions In this study population, a questionnaire with 7-day recall provided information consistent with a daily diary measure of the average week-long experience of T2D symptoms and impacts.

Keywords Questionnaires · Mental recall · Validation studies · Diabetes mellitus type 2 · Signs and symptoms

Abbreviations

CCC	Concordance correlation coefficient
DSC-R	Diabetes symptom checklist-revised
FDA	Food and Drug Administration
HbA1c	Glycated hemoglobin
PRO	Patient-reported outcome
SF-36v2 Acute	Short-form 36 health survey, version 2 Acute

Introduction

Patient-reported outcome (PRO) instruments as standard tools of evaluation have been a significant step toward assuring patient-centered health care. Patients' perception of their own change in health and function can differ from that of health care professionals, and patient experience is important in evaluating the treatment of chronic diseases [1]. Selecting an appropriate recall period and providing justification for it is one task in the process of developing a PRO [2]. The optimal recall period depends in part on the characteristics of the event being measured, including the frequency of occurrence or rate of fluctuation [3]. Psychological theory about cognitive processes of memory suggests

A. V. Bennett (✉)
Health Outcomes Research Group, Memorial Sloan-Kettering
Cancer Center, New York, NY, USA
e-mail: bennetta@mskcc.org

D. L. Patrick · P. Diehr
Department of Health Services, University of Washington,
Seattle, WA, USA

D. M. Bushnell
Health Research Associates, Inc, Mountlake Terrace, WA, USA

C. F. Chiou
Amgen, Inc, Thousand Oaks, CA, USA

P. Diehr
Department of Biostatistics, University of Washington,
Seattle, WA, USA

that retrospective recall is likely not accurate [e.g., 4]; however, little empirical evidence exists to either support or discredit the use of particular recall periods in patient-reported outcomes. The goal of this analysis was to compare daily diary and weekly measurement to determine whether they provide similar information about the week-long symptom experience of patients with type 2 diabetes.

The accuracy of short-term retrospective recall, i.e., recall over a period of 1–4 weeks, in measuring daily events has been tested in urinary incontinence [5–7], physical activity [8, 9], and alcohol consumption [10, 11]. The correlation between daily diary and retrospective reporting in these studies varied from 0.33 to 0.89. In a literature review of recall bias in alcohol consumption, studies are presented in support of each possible finding: over-reporting, under-reporting, and no difference in retrospective measurement compared to daily diary records [11]. A significant methodological shortcoming of these studies is that findings are based on Pearson's correlation coefficient. A correlation coefficient alone does not incorporate differences in mean scores or the slope of the relationship between each type of measurement (e.g., if the two types of measurement were the same, the means would be equal and the slope would be 45°).

Three studies of recall bias in the measurement of pain compared retrospective reports to repeated real-time reporting i.e., ecological momentary report. This work found (1) recalled pain is most similar to the maximum and most recent real-time 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) although group-level correspondence between real-time and retrospective reports is moderate, the within-person correspondence is low [12–14].

These few existing studies on short-term retrospective recall suggest that there is difference between retrospective report and daily experience, and that the amount of difference depends on the patterns of daily responses such as the amount of day-to-day variation in daily responses. The symptoms studied thus far are only incontinence and pain [5–7, 12–14].

The study reported here compared daily diary and weekly questionnaire responses to items measuring symptoms and impacts of type 2 diabetes. The analysis is organized into three parts: (1) identifying any effect of repeated daily measurement on weekly item response, to determine whether the study design is suitable for this analysis; (2) exploration of the relationship between weekly and daily diary reporting, including the comparison of group means and the concordance of scores; and (3) comparison of the ability of weekly retrospective and daily measurement to detect differences between known groups and to provide preliminary evidence of whether one recall period may be more likely to detect change in a treatment trial. Within Part 2 of the analysis, two hypotheses based on cognitive theory of memory and findings from studies of pain recall were tested. First, the weekly retrospective response was expected to be most similar to the maximum daily report and the most recent daily report, in this case, Day 7. Second, the difference between the weekly and mean daily responses was expected to be larger in items with greater variability in daily responses.

Methods

Study population

The diary data were collected during an observational study in 144 adult patients with type 2 diabetes whose HbA1c was $\geq$6.0. Subjects were recruited as a de novo convenience sample through 5 clinic sites located in multiple regions of the United States. Subjects were excluded if they had recently changed their medication because it could make their symptoms much more unstable than usual. Subjects were also excluded if they had clinical depression or cognitive impairment because it could affect their memory.

Study design

All subjects provided informed consent prior to enrollment. Within each of the 5 study sites, participants were randomized in equal numbers to one of two groups (explained below) after phone screening and before the first study visit. The study period was 14 days long. In Week 1, all subjects completed the daily diary each day and the weekly questionnaire on Day 7. In Week 2, half the subjects (Group 1) did not complete the daily diary and half the subjects (Group 2) did complete the daily diary each day, and all subjects completed the weekly questionnaire on Day 14. All subjects completed the SF-36v2 Acute on Day 7 and Day 14. Subjects were instructed to complete the daily diary and questionnaires 2 h after dinner.

Data collection instruments

Daily diary and weekly questionnaire

The daily diary comprised 7 identical daily diary cards, with 24-h recall, that measured symptoms and impacts experienced by people with type 2 diabetes. The 9 items from the diary included in this analysis measured (1) frequency of thirst, (2) difficulty keeping blood sugar in a recommended range, (3) difficulty concentrating, (4) difficulty getting along with others, (5) difficulty doing daily tasks, (6) irritableness at its worst, (7) tiredness at its worst, (8) worst tingling in feet, and (9) worst muscle aches and pains. For each item, a response of "0" indicated the symptom or impact was not experienced. Items 1 and 2 were measured on a 5-point scale (0–4) where "4" indicates "All the time" or "A great deal". Items 3 through 9 were measured on an 11-point scale (0–10) where "10" indicates the symptom or impact was as difficult or extreme as the subject could imagine. Only diaries with at least 5 of 7 days completed were included in analysis. The weekly questionnaire was a replication of a daily diary card, except the recall period was 7 days instead of 24 h.

The diary items were developed per the process described in the 2009 FDA Guidance for Industry [2], including concept elicitation interviews, cognitive interviewing, and analysis of psychometric properties. Individual items from the daily diary were tested for psychometric validity in a separate analysis of the same diary data described in this paper. Construct validity was confirmed via correlation (or lack thereof) with items and subscales of the Diabetes Symptom Checklist-Revised (DSC-R) [15] and the SF-36v2 Acute. Internal consistency was not assessed because items scores were not intended to be combined into sub-scales scores or a total score. Test–retest reliability was confirmed (Spearman-Brown range 0.72–0.93). Although the concept elicitation and cognitive interviews supported the relevance of each item, analysis of the diary data indicated the low prevalence of some items in this population. In the Week 1 diary, 49% of patients never reported any difficulty getting along with others (item #4), 40% never reported any difficulty doing daily tasks (item #5), and 45% never reported any tingling in feet (item #8). The other 5 symptoms and impacts were each experienced by 64–96% of patients. Table 3 indicates missing data, which were low, and the variability of item responses.

SF-36v2 Acute

The SF-36 Acute, 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 10-item Physical Functioning subscale and 5-item Mental Health subscale scores were incorporated in this analysis and are scored on a scale of 0–100, in which higher scores indicate better health [16, 17].

Date–time stamp

An electronic, preprogrammed date and time stamp (DYMO® DateMark Date/Time Stamper; DYMO®, Stamford, CT) was provided to subjects to document the exact date and time they completed the daily diary, the weekly questionnaire, and the SF-36v2 Acute. Subjects were not able to alter the date or time of the preprogrammed stamp.

Patient health and demographic information

The patient's most recent (in past 30 days) test result for percentage of glycated hemoglobin (HbA1c), date of diagnosis, and treatment regimen was provided by the clinic. Patients with HbA1c $\leq$ 7.0 are considered to have blood sugar in tight control and have lower risk of complications [18]. Subjects were asked to report their date of birth, sex, ethnic group, education level (in years), marital status, employment status, and general health.

Human subjects research

Institutional review board approvals were granted for collection of the data and for secondary use of the data in this analysis; data collection and analysis were conducted in accordance with ethical standards described in the 1964 Declaration of Helsinki.

Data analyses

In each analysis described in this paper, the 9 items were analyzed individually and all p-values reported are based on two-sided 95% confidence intervals, unless otherwise noted. The response scales were treated as continuous variables.

Identifying any effect of repeated daily measurement on weekly item response

The primary analyses reported in this paper were conducted on the Week 1 study data, in which all subjects (Group 1 and Group 2) completed the daily diary on Days 1–7 and the weekly questionnaire on Day 7. If completing the daily diary affected responses to the weekly questionnaire, for instance by making subjects more aware of their daily symptoms, then results from the primary analyses could not be generalizable to a study setting where study subjects do not complete a daily diary. To identify whether repeated daily measurement had an effect on the weekly item response, the Week 2 weekly responses of patients completing the daily diary (Group 2) were compared using t-tests to the Week 2 weekly responses of patients who did not complete the daily diary during the recall period of the weekly measurement (Group 1).

Exploration of the relationship between weekly and daily diary reporting

To test the first hypothesis—that the weekly report would be most similar to the maximum daily report and the most recent daily report (Day 7), for each item, the weekly item response was compared with summary descriptors of the daily diary item responses (e.g., mean of daily responses, median of daily responses, maximum daily response, last day) to identify what the weekly item response most closely approximated. A list of each descriptor and what it indicates is shown in Table 1. The weekly response was compared with each descriptor of the daily diary response by (a) dependent sample t-tests for comparison of group means and (b) a measure of concordance. The amount of difference in group means was evaluated by whether it was less than 0.5 standard deviations. To calculate this effect size, the difference was divided by the average of the two standard deviations ($SD_{\text{mean daily diary}}$ and $SD_{\text{weekly response}}$). Concordance was determined by estimating the concordance correlation coefficient (CCC), a statistic composed of Pearson's correlation coefficient and a bias correction factor calculated from the differences in the means and variance of the two variables. The bias correction factor indicates the amount of deviation of the Pearson's best-fit line from the 45° origin line (the concordance line); possible values range from 0 to 1 where values closer to 1 indicate less deviation. The CCC ranges from -1 to $+1$, where values closer to 1 indicate greater concordance [19, 20].

The CCC was interpreted per the threshold of agreement recommended for the intra-class correlation coefficient (ICC) because the statistics are algebraically similar and consistent with each other [20]. A one-sided ($\alpha = 0.05$) *t* test was conducted for each item to determine whether the CCC for the weekly response and mean of daily reports is $\geq$0.70, indicating a high level of agreement.

To test the second hypothesis—that the difference between the weekly item response and mean of daily item responses would be larger in cases where there was greater variability in daily reports, the difference was analyzed by the standard deviation of the 7 daily reports. The difference between the weekly item response and mean of daily item responses was regressed on to the standard deviation of daily reports to test for a linear trend. Additionally, for each item, the difference was regressed onto the range of daily reports. The range was defined as the maximum minus minimum daily report.

Comparison of the ability of weekly retrospective and daily measurement to detect differences between known groups

The ability of each type of measurement to detect differences between known groups (i.e., discriminant ability) was compared by calculating and comparing *t* statistics for the differences in item scores by sample subgroups for each type of measurement. The subgroups were defined by low and high tertiles of SF-36 Physical Functioning subscale scores, by low and high tertiles of SF-36 Mental Health subscale scores, and by low and high levels of HbA1c (6.0–6.9 vs. 8.0–8.9). The null hypothesis of each *t* test was that there was no difference between sample subgroups.

Results

Patient characteristics

Subjects were recruited and enrolled through 5 clinic sites located across the United States. Of the 144 subjects enrolled, 142 (98.6%) completed the study (two subjects withdrew from the study after 1 and 3 days). An additional

Table 1 Summary descriptors of the daily diary item responses

Descriptor	What it indicates	Difference[a]		Concordance[b]	
		Median	Range	Median	Range
Mean	Average of daily reports	0.39	0.16–0.77	0.82	0.65–0.86
Median	Median of daily reports	0.53	0.16–0.95	0.79	0.55–0.85
Mode, max	Most common response (largest mode if multiple modes)	0.47	0.14–0.86	0.77	0.57–0.82
Mode, min	Most common response (smallest mode if multiple modes)	0.69	0.17–1.12	0.73	0.50–0.79
Maximum	Most severe response	−0.72	−1.22–−0.32	0.71	0.60–0.87
Minimum	Least severe response	1.04	0.65–2.04	0.50	0.25–0.64
Day 7 (last day)	Response of the same day the weekly report	0.41	0.22–0.93	0.76	0.55–0.80
Day 6	Response of the day before the weekly report	0.54	0.19–0.93	0.72	0.53–0.79
Day 1 (first day)	First daily report	0.08	−0.07–0.30	0.67	0.52–0.84

One hundred and forty patients completed the 9 items in both the daily diary and weekly questionnaire

[a] The *difference* is the weekly report minus the descriptor (shown at left) for each of the 9 questionnaire items; it is summarized in this table by the median and range of values for the 9 items. E.g., the difference between the weekly report and the mean of daily reports ranged from 0.16 to 0.77 across the 9 items, with a median value of 0.39

[b] The *concordance* is the concordance correlation coefficient for the weekly report with the descriptor (shown at left) for each of the 9 questionnaire items. It is also summarized here by the median and range of values for the 9 items

2 were removed from the analysis (one for non-compliance and one for investigator withdrawal as cognitively impaired), leaving 140 (97.2%) in the sample. Analyses (not shown) indicated that the randomization was effective. It created two groups that are not significantly different in all but 5 (about 1%) of the approximately 350 variables collected in the observational study.

The mean (SD) age of the sample ($n = 140$) was 62.2 (11.4) years; 57% were men and 78% were White (non-Hispanic). The mean (SD) HbA1c% was 7.7 (1.3) with an average length of time since diabetes diagnosis of 12.6 (8.4) years. Sixty-four percent were married, 51% were actively employed, and 40% were retired. The majority of the sample (88%) indicated having good to excellent general health. Table 2 shows patient characteristics in detail.

Diary response rates

The number of missing entries out of a total 980 (7 days × 140 patients) entries for each item ranged from 11 (1.1%) to 17 (1.7%). The number of patients that completed at least 5 out of 7 days of a diary item ranged from 136 to 139 across items (see Table 3). Analysis of the date–time stamps on each patient's diary indicate that 77.9% (109/140) of patients stamped their daily diary at the correct date and time on at least 5 out of 7 days. Specifically, the diary was stamped correctly for 7 out of 7 days by 42.1% (59/140) of patients, for 6 out of 7 days by 22.1% (31/140) of patients, and for 5 out of 7 days by 13.6% (19/140) of patients.

Diary characteristics

For each item, the study sample's average weekly response and mean of daily responses is reported in Table 3. The average within-person variation in daily item responses, i.e., the standard deviation of a patient's daily responses to an item, ranged from 0.41 to 0.60 for items with a 5-point scale and from 0.51 to 1.23 for items with an 11-point scale.

Analyses

Identifying any effect of repeated daily measurement on weekly item response

The differences in the Week 2 weekly questionnaire responses of Group 1 and Group 2 were not statistically significant for any items at $P < 0.05$ (not shown), suggesting that completing the daily diary in the 7 days preceding the weekly report had no measurable effect on the weekly report.

Table 2 Patient characteristics

N	140
Age *mean (SD, range)*	62.2 (11.4, 33–91)
Gender *n* (%)	
Male	80 (57.1%)
Female	60 (42.9%)
HbA1c *mean (SD, range)*	7.7 (1.3, 6–12.9)
Years since diagnosis *mean (SD, range)*	12.6 (8.4, 1–59)
Treatment *n* (%)	
Oral agents only	71 (50.7%)
Oral agents and insulin	43 (30.7%)
Insulin only	19 (13.6%)
Byetta	7 (5.0%)
Ethnic group *n* (%)	
White (non-Hispanic)	109 (77.9%)
Black/African American	15 (10.7%)
Asian/Pacific Islander	7 (5.0%)
Hispanic	8 (5.7%)
Other: Jamaican	1 (0.7%)
Marital status *n* (%)	
Married	89 (63.6%)
Widowed	15 (10.7%)
Separated	0 (0.0%)
Divorced	22 (15.7%)
Never married	14 (10.0%)
Employment *n* (%)	
Employed	71 (50.7%)
Retired	56 (40.0%)
Homemaker	2 (1.4%)
Unemployed	9 (6.4%)
Missing	2 (1.4%)
Highest grade completed ($n = 137$) *mean (SD, range)*	14.1 (2.5, 8–20)
General health *n* (%)	
Excellent	6 (4.3%)
Very good	29 (20.7%)
Good	74 (52.9%)
Fair	28 (20.0%)
Poor	2 (1.4%)
Missing	1 (0.7%)
SF-36 Physical Functioning subscale *mean (SD, range)*	44.1 (10.9, 17.0–57.0)
SF-36 Mental Health subscale *mean (SD, range)*	52.4 (9.1, 19.0–64.1)

Exploration of the relationship between weekly and daily diary reporting

The comparison of the weekly response to each summary descriptor of the daily responses via group means showed consistent trends among the 9 items. For each item, the weekly response was larger than the mean of daily

Table 3 Daily diary and weekly item responses

Item	Measure	*N*	Sample mean	Sample SD	Within-person SD	Weekly minus mean of daily	ES	CCC	Rho	Bias factor
1. Thirsty (0–4)	Weekly	137	1.60	0.76						
	Mean of daily	137	1.43	0.76	0.41	0.16*	0.21	0.83†	0.85	0.98
2. Difficulty keeping blood sugar in rec. range (0–4)	Weekly	136	2.13	0.99						
	Mean of daily	136	1.95	0.94	0.60	0.18*	0.19	0.86†	0.88	0.98
3. Difficulty concentrating (0–10)	Weekly	136	1.76	1.94						
	Mean of daily	136	1.42	1.57	0.83	0.36*	0.21	0.82†	0.85	0.96
4. Difficulty getting along with others (0–10)	Weekly	136	1.18	1.69						
	Mean of daily	136	0.79	1.25	0.59	0.39*	0.27	0.78†	0.85	0.92
5. Difficulty doing daily tasks (0–10)	Weekly	136	1.55	1.91						
	Mean of daily	136	1.34	1.64	0.75	0.22**	0.12	0.85†	0.86	0.98
6. Irritable (0–10)	Weekly	138	2.27	2.26						
	Mean of daily	138	1.50	1.55	0.99	0.77*	0.40	0.65	0.75	0.87
7. Tired (0–10)	Weekly	139	3.48	2.52						
	Mean of daily	139	2.84	1.88	1.23	0.64*	0.29	0.75	0.82	0.92
8. Tingling in your feet (0–10)	Weekly	139	1.69	2.31						
	Mean of daily	139	1.25	1.90	0.51	0.45*	0.21	0.85†	0.89	0.96
9. Muscle aches and pains (0–10)	Weekly	139	2.60	2.32						
	Mean of daily	139	2.14	2.04	0.79	0.47*	0.22	0.80†	0.83	0.97

Scores of zero indicate symptom or impact is not experienced

ES Effect size of the difference between weekly reports and the mean of daily reports, *CCC* Concordance correlation coefficient, *Rho* Pearson's correlation coefficient, *Bias factor* Bias adjustment factor of the CCC

Weekly minus mean of daily reports is non-zero: * $P < 0.001$, ** $P < 0.01$

† CCC is ≥0.70 ($P < 0.05$)

responses ($P < 0.001$; for item 5: *getting along with others* $P < 0.01$) and was smaller than the maximum of daily responses. The difference between the weekly and mean of daily responses ranged from 0.16 to 0.18 for items with a 5-point scale and from 0.22 to 0.77 for items on an 11-point scale. For each item, the difference between the weekly and mean of daily responses was less than 0.5 standard deviations. The difference between the weekly and maximum of daily responses ranged from 0.32 to 0.61 for items with a 5-point scale and from 0.36 to 1.23 for items on an 11-point scale; this difference was significant ($P < 0.05$) for all items except Item 8: *Tingling in feet*. The difference between the weekly and last daily response (Day 7) ranged from 0.21 to 0.25 for items with a 5-point scale ($P < 0.001$) and from 0.25 to 0.93 for items with an 11-point scale ($P < 0.01$); for 8 of 9 items, this difference was larger than the difference between the weekly and mean daily report.

The weekly response had the highest concordance with the mean of daily responses compared to other descriptors in 6 of the 9 items. For the remaining 3 items, the concordance was highest with the maximum daily report. In the comparison of the weekly and mean of daily reports, the concordance statistic ranged from 0.65 to 0.86 and for each item, it was lower than the correlation statistic (range 0.75–0.89) because the bias correction factor was less than 1.00 (range 0.87–0.98). For 7 of 9 items, the CCC was greater than or equal to 0.70 ($P < 0.05$), indicating an acceptable level of agreement. The concordance of the weekly and maximum report ranged from 0.55 to 0.80. The range of difference and concordance for each summary descriptor is shown in Table 1.

The difference between the weekly response and the mean of daily responses varied by the standard deviation of daily responses for four items; these items are Item 1: *Thirsty* ($P < 0.01$), Item 4: *Getting along with others* ($P < 0.01$), Item 6: *Irritable* ($P < 0.001$), and Item 8: *Tingling in your feet* ($P < 0.001$). For these four items, the difference was larger in observations with a larger standard deviation of daily reports. Regression of the difference onto the range of daily responses gave the same results.

Comparison of the ability of each type of measurement to detect differences between known groups

The mean of daily responses and the weekly report were compared by the statistical significance of the amount of difference in the item scores between subjects with high versus low SF-36 Physical Functioning subscale scores, SF-36 Mental Health subscale scores, and HbA1c values. Across the 3 patient characteristics and 9 items, the significance of the t-statistics was the same for the mean daily and weekly report in 23 (85%) of the 27 comparisons (Table 4). Of the 4 (15%) comparisons in which the significance of the two types of measurement was not the same, only the mean of daily reports was significant in 2 comparisons and only the weekly report was significant in the other 2. The absolute value of the *t* statistic was larger for the weekly report in 16 (59%) of the 27 comparisons.

Discussion

Two hypotheses were tested in this analysis. The first hypothesis was the weekly response would be most similar to the maximum daily report and the last daily report, in this case, Day 7; this hypothesis was not confirmed. This analysis found the weekly report was consistently higher than the mean of their daily reports, and the difference between the weekly report and the mean of daily reports was less than 0.5 standard deviations. When the weekly report was compared to summary descriptors of the daily reports (mean, median, maximum, etc.), the weekly report was most concordant with the mean of daily reports, and

Table 4 Comparison of discriminant ability of mean of daily responses and weekly report

	Item	Mean of daily responses			Weekly report		
		Difference in scores of high and low groups	*t*	*P*	Difference in scores of high and low groups	*t*	*P*
PF subscale	*1	−0.45	−2.516	0.014	−0.52	**−2.879**	0.005
	*2	−0.03	**−0.158**	0.875	0.00	0.000	1.000
Groups	*3	−1.25	−3.653	0.000	−1.47	**−3.875**	0.000
High: 51+	*4	−0.44	**−1.392**	0.168	−0.35	−0.897	0.372
Low: <40	*5	−1.89	**−5.086**	0.000	−1.83	−4.628	0.000
	6	−0.84	**−2.357**	0.021	−0.79	−1.561	0.123
	*7	−1.53	−3.585	0.001	−2.01	**−3.745**	0.000
	*8	−0.96	−2.193	0.031	−1.28	**−2.532**	0.013
	*9	−1.90	**−4.175**	0.000	−1.98	−4.065	0.000
MH subscale	1	−0.25	−1.397	0.166	−0.41	**−2.448**	0.016
	*2	−0.34	−1.633	0.106	−0.41	**−1.874**	0.064
Groups	*3	−2.03	**−6.209**	0.000	−2.21	−5.551	0.000
High: 58+	*4	−1.14	−3.759	0.000	−1.70	**−4.609**	0.000
Low: <50	*5	−1.99	−5.277	0.000	−2.36	**−5.421**	0.000
	*6	−1.71	−5.021	0.000	−2.41	**−5.045**	0.000
	*7	−2.01	−4.706	0.000	−2.67	**−5.098**	0.000
	8	−0.85	−1.757	0.082	−1.40	**−2.472**	0.015
	*9	−1.84	−3.934	0.000	−2.56	**−5.134**	0.000
%HbA1c	*1	0.18	0.818	0.416	0.18	**0.835**	0.407
	2	0.57	**2.835**	0.006	0.36	1.588	0.117
Groups	*3	0.21	0.473	0.637	0.24	**0.483**	0.631
High: 8.0–8.9	*4	0.10	**0.250**	0.803	0.01	0.023	0.982
Low: 6.0–6.9	*5	0.58	**1.036**	0.304	0.38	0.621	0.537
	*6	−0.01	−0.030	0.976	0.53	**0.779**	0.439
	*7	0.07	0.152	0.879	0.42	**0.672**	0.504
	*8	1.11	**1.812**	0.074	0.88	1.269	0.209
	*9	1.12	**1.963**	0.054	1.05	1.745	0.085

The *t*- statistic with the larger absolute value is shown in bold

PF SF-36 Physical Functioning subscale, *MH* SF-36 Mental Health subscale; both subscales are scored on a 0–100 scale where higher scores indicate better physical functioning or mental health

* Comparisons in which the significance of the *P* values are the same

the difference between the weekly report and the mean of daily reports was smaller than the difference between the weekly report and other descriptors of the daily reports. These results support that although the weekly report was slightly higher than the mean of the daily reports, weekly reports of the study subjects were most similar to an average of their daily symptom experience during the recall period.

The second hypothesis was the difference between weekly and mean daily responses would be larger with greater variability in daily responses; this hypothesis was confirmed in only some items. In four of the nine items, the amount of variation in the daily reports was associated with the difference between weekly and the mean of daily reports. It was expected that greater variability in daily reports would be associated with a larger difference. It is possible that this association was not found in some items because of the limited amount of variation in the daily reports in this study sample.

Strengths of this study included that the weekly report questionnaire and daily diary cards were identical in wording and layout except for the recall period. The design of the data collection also made it possible to test whether the weekly report was affected by subjects having completed the daily diary. The comparison of the Week 2 weekly responses indicates the weekly report was not affected by completing the daily diary during the recall period. It is not known with certainty whether completing the daily diary in Week 1 would make patients more aware of their symptoms and impacts during Week 2 to the point that it affected their responses to the Week 2 weekly report. A limitation of this study is the use of paper diaries with an electronic date–time stamp instead of an entirely electronic diary. Review of the date–time stamp data indicated 77.9% of patients completed 5 of 7 diaries (in Week 1) at the correct date and time.

The weekly report and the mean of daily reports were highly concordant, and the average difference between weekly and mean of daily reports was small. Although the difference varied more widely for individuals, either of these two types of measurement could be used to estimate the average experience of a group of patients during a 7-day period. Only a daily diary would be able to detect day-to-day variation in symptoms and impacts.

The extra burden on study participants of completing the diary each day compared to a once a week report, and the additional study resources required for collecting and analyzing diary data are practical considerations in choosing between daily diary and weekly measurement of patient experience. The weekly report and the mean of daily reports were equally sensitive to detecting differences by item score between groups of patients defined by health status. With this study data, it was not possible to compare the sensitivity of each type of measurement to change over time.

The findings of this study are based on cross-sectional data from a study sample of adults with T2D whose average time since diagnosis was 12.6 years. Although the sample was stratified by disease severity, specifically by level of HbA1c, the prevalence of symptoms that are noticeable to patients is also dependent on the natural history of disease. Most items in this analysis asked patients to report their worst experience of the symptom or impact during the recall period, (e.g., worst difficulty concentrating or worst muscle aches and pains). These characteristics of the study sample and questionnaire items may limit the generalizability of results. However, similar relationships were found when the same analyses were conducted on patient reporting of respiratory symptoms in separate studies of cystic fibrosis [21] and chronic obstructive pulmonary disease (manuscript in progress).

The results of the analysis in this sample of T2D patients support that there is consistency in a group of patients' weekly retrospective reports with the average of their reported daily experience. Choice of daily diary and/or weekly recall of T2D symptoms may well depend on the desired sample criteria, such as newly diagnosed patients, frequency of assessment, study design, and other important characteristics of studies that interact with reports from patients using varying length of recall.

Acknowledgments We would like to thank Mona Martin RN MPA, Health Research Associates, Inc, for guidance in data collection, Shravanthi Gandra PhD MBA, Amgen, Inc, for her early participation in this study, and Denise Globe PhD and Xing-yue Huang PhD, Amgen Inc, for their contributions to planning the data collection. The data were made available through a study with multiple aims sponsored by Amgen, Inc. in a contract with Health Research Associates, Inc and Donald Patrick. Chiun-Fang Chiou is an employee of Amgen.

References

1. Revicki, D. A., Osoba, D., Fairclough, D., Barofsky, I., Berzon, R., Leidy, N. K., et al. (2000). Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Quality of Life Research, 9*(8), 887–900.
2. FDA. (2009). Patient reported outcome (PRO) measures: Use in medical product development to support labeling claims. *Federal Register, 74*(235), 65132–65133.
3. Stull, D. E., Leidy, N. K., Parasuraman, B., & Chassany, O. (2009). Optimal recall periods for patient-reported outcomes: Challenges and potential solutions. *Current Medical Research and Opinion, 25*(4), 929–942.
4. Hufford, M. R., & Shiffman, S. (2003). Assessment methods for patient-reported outcomes. *Disease Management and Health Outcomes, 11*(2), 77–86.
5. Homma, Y., Ando, T., Yoshida, M., Kageyama, S., Takei, M., Kimoto, K., et al. (2002). Voiding and incontinence frequencies:

Variability of diary data and required diary length. *Neurourology and Urodynamics, 21*(3), 204–209.

6. Kenton, K., Fitzgerald, M. P., & Brubaker, L. (2006). What is a clinician to do-believe the patient or her urinary diary? *Journal of Urology, 176*(2), 633–635. (discussion 5).
7. Elser, D. M., Fantl, J. A., & McClish, D. K. (1995). Comparison of "subjective" and "objective" measures of severity of urinary incontinence in women. Program for Women Research Group. *Neurourology and Urodynamics, 14*(4), 311–316.
8. Richardson, M. T., Ainsworth, B. E., Jacobs, D. R., & Leon, A. S. (2001). Validation of the Stanford 7-day recall to assess habitual physical activity. *Annals of Epidemiology, 11*(2), 145–153.
9. Stel, V. S., Smit, J. H., Pluijm, S. M., Visser, M., Deeg, D. J., & Lips, P. (2004). Comparison of the LASA Physical Activity Questionnaire with a 7-day diary and pedometer. *Journal of Clinical Epidemiology, 57*(3), 252–258.
10. Gmel, G., & Daeppen, J. B. (2007). Recall bias for seven-day recall measurement of alcohol consumption among emergency department patients: Implications for case-crossover designs. *Journal of Studies on Alcohol and Drugs, 68*(2), 303–310.
11. Hilton, M. E. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction, 84*(9), 1085–1092.
12. Stone, A. A., Broderick, J. E., Kaell, A. T., DelesPaul, P. A., & Porter, L. E. (2000). Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritics? *The Journal of Pain, 1*(3), 212–217.
13. Stone, A. A., Broderick, J. E., Shiffman, S. S., & Schwartz, J. E. (2004). Understanding recall of weekly pain from a momentary assessment perspective: Absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain, 107*(1–2), 61–69.
14. Stone, A. A., Schwartz, J. E., Broderick, J. E., & Shiffman, S. S. (2005). Variability of momentary pain predicts recall of weekly pain: A consequence of the peak (or salience) memory heuristic. *Personality and Social Psychology Bulletin, 31*(10), 1340–1346.
15. Grootenhuis, P. A., Snoek, F. J., Heine, R. J., & Bouter, L. M. (1994). Development of a type 2 diabetes symptom checklist: A measure of symptom severity. *Diabetic Medicine, 11*(3), 253–261.
16. Ware, J. E., Jr., Kosinski, M., & Dewey, J. E. (2000). *How to score version 2 of the SF-36 health survey*. Lincoln, RI: Quality Metric, Incorporated.
17. Ware, J. E, Jr., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health survey manual and interpretation guide*. Boston, MA: New England Medical Center, The Health Institute.
18. American Diabetes Association. (2009). Standards of medical care in diabetes–2009. *Diabetes Care, 32*(Suppl 1), S13–S61.
19. Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268.
20. Deyo, R. A., Diehr, P., & Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clinical Trials, 12*(4 Suppl), 142S–158S.
21. Bennett, A. V., Patrick, D. L., Lymp, J. F., Edwards, T. C., & Goss, C. H. (2010). Comparison of 7-day and repeated 24-hour recall of symptoms of cystic fibrosis (in press). doi:10.1016/j.jcf.2010.08.008.