BMJ Open

Recall of health-related quality of life: how does memory affect the SF-6D in patients with psoriasis or multiple sclerosis? A prospective observational study in Germany

Janine Topp ,[1] Valerie Andrees,[1] Christoph Heesen,[2] Matthias Augustin,[1] Christine Blome[1]

To cite: Topp J, Andrees V, Heesen C, *et al*. Recall of health-related quality of life: how does memory affect the SF-6D in patients with psoriasis or multiple sclerosis? A prospective observational study in Germany. *BMJ Open* 2019;**9**:e032859. doi:10.1136/bmjopen-2019-032859

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-032859).

Received 09 July 2019
Revised 25 September 2019
Accepted 25 October 2019

[1]Institute for Health Services Research in Dermatology and Nursing (IVDP), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany
[2]Department of Neurology, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany

Correspondence to
Janine Topp; j.topp@uke.de

ABSTRACT

Objective This study aimed to quantify recall bias in the measurement of health-related quality of life (HRQoL), that is, the extent to which recollection is impaired and leads to distorted judgements.

Design Prospective observational study.

Setting and participants One hundred patients with two paradigmatic chronic diseases (50 with multiple sclerosis and 50 with psoriasis) were recruited at two outpatient clinics.

Methods and outcome measures Patients completed the online version of the 12-Item Short Form Survey (SF-12) repeatedly for 28 consecutive days: (1) daily, considering the past 24 hours; (2) weekly, considering the past 7 days; and (3) on the last day of data collection, considering the past 4 weeks. SF-12 scores for all three measurement approaches were subsequently converted into preference-based utility indices (Short-Form Six-Dimension). Agreement of the three indices was analysed on group and individual patient levels.

Results The mean age of participants was 40.3 years (±12.0), and 63% were female. The utility index based on daily recall (0.74±0.13) was more positive than indices based on a weekly (0.70±0.13, p<0.001) or a monthly (0.70±0.14, p<0.001) recall. While agreement of measurement approaches was high on group level (intraclass correlation coefficient>0.85), it was lower for the subgroup of patients experiencing high variability of HRQoL over time. Bland-Altman plots revealed considerable differences on individual patient level.

Conclusions On the group level, retrospective overestimation and underestimation of HRQoL almost cancelled out one another and recall bias was relatively small. Therefore, a 4-week recall period could be appropriate when group-level data are used for research or economic evaluations. In contrast, recall bias can be considerable on the individual patient level and may thus impact decision-making in clinical practice.

Trial registration number VfD_RECALL_16_003837.

Strengths and limitations of this study

- The study design allows for direct intraindividual comparisons between retrospective and near real-time reporting of health-related quality of life (HRQoL).
- In contrast to paper-based diaries, the repeated data collection via online questionnaires reduces the number of missing values and facilitates monitoring of the time of data entry.
- A validated questionnaire that is very frequently used in research and economic evaluations was used to analyse recall bias in HRQoL.
- A convenience sample of patients diagnosed with multiple sclerosis or psoriasis was recruited, and generalisability of findings might thus be limited.
- Participants completed both questionnaires with a recall period of 1 day and questionnaires with a recall period of 1 or 4 weeks; daily completion may have improved the week and month recall so that recall bias may be underestimated due to the specific study design.

INTRODUCTION

Measuring health-related quality of life (HRQoL) is by no means a simple task. The underlying construct is complex, subjective and not directly observable.[1] The widely accepted strategy to approach the construct is to ask patients about their perceived HRQoL using standardised surveys. Comprising questions are assumed to reflect important domains of HRQoL, such as physical and social functioning or mental health. Subsequently, HRQoL reports are used to assist decision-making and monitoring in clinical practice, to assess the effectiveness of interventions in clinical trials and to determine treatment benefit in economic evaluations.[2–4]

For economic evaluations, HRQoL reports of patients are weighted according to predetermined preferences, which reflect the value that people place on the various domains of HRQoL. The resulting utility values are

used to estimate quality-adjusted life years (QALYs), an important component of many economic evaluations.[5] Thus, utility values are of great significance when weighing up costs and benefits of a new treatment and can inform the decision as to whether reimbursement of treatment costs are recommended.[6 7]

Many HRQoL questionnaires refer to a specific retrospective period, asking patients to recall their impairment during the past day, the past week or the past month.[8–11] In general, the ability to remember previous states influences how accurately patients report their HRQoL. The longer the recall period, the higher the probability of recall bias. Recall bias, also called memory bias, is understood as the extent to which memory is limited, leading to distorted judgements of the target construct.[12] Hence, the ability to accurately remember and report HRQoL affects reliability and validity of the used instrument.

Recall bias is not unique to HRQoL assessment but has already been observed for self-reports on health-related events, health behaviours and symptoms.[8] Research on patients' ability to recall pain, for example, indicates a retrospective overestimation of symptom severity.[13 14] The association between diary data and retrospective data was found to be moderate only.[15] Additionally, retrospective pain ratings are disproportionally affected by the most recent and the highest pain levels within the recall period (peak-end effect).[16 17] Consequently, a peak-end effect could also impact retrospective HRQoL assessment.[18] In addition, little is known about the impact of HRQoL fluctuations on the ability to recall HRQoL states.

An assessment of the past day, that is, a short recall period, reduces the risk of recall bias. Conversely, a 1-day report is accompanied by information loss and limits generalisability because overall HRQoL of a patient with a chronic disease could substantially differ from day to day.[8 19] The stated trade-off between generalisability on the one hand and recall bias on the other hand emphasise the difficulty in determining the optimal recall period and defining a universal standard for HRQoL assessment.

For this reason, some HRQoL surveys are available in different versions referring to different recall periods.[9] This applies, for example, to the Short-Form Six-Dimension (SF-6D) health index,[20] a preference-based utility estimate that can be calculated based on different versions of the 12-Item Short Form Survey (SF-12): next to the standard version referring to the HRQoL of the past 4 weeks, an acute (ie, past week) version and a daily (ie, past 24 hours) version are available. In the present study, recall bias is assumed when repeated assessment on a daily basis and retrospective assessment of the same period of time do not agree with one another.

We investigated recall bias in a group of chronically ill individuals, including patients diagnosed with psoriasis or multiple sclerosis (MS). Both diseases are associated with significant impairments in HRQoL,[21 22] and maintaining or improving HRQoL is an important treatment goal. This emphasises the need for reliable and valid measurement instruments for clinical practice, research and economic evaluations.

The main objective of this study was to assess the agreement of preference-based HRQoL reports with different recall periods gathered over a period of 4 weeks. Averaged daily reports, averaged weekly reports and a retrospective report over the entire 4-week period were compared. We further explored whether the agreement of HRQoL reports with different recall periods is affected by observed dynamics in daily reports.

METHODS

Setting and participants

We conducted a longitudinal observational study and followed the reporting guideline for observational studies in epidemiology (Strengthening the Reporting of Observational Studies in Epidemiology statement).[23] Patients were recruited through the outpatient clinics for MS or psoriasis. Patients were eligible to participate in the study if they were diagnosed with psoriasis or MS, were at least 18 years of age, and had internet access and an email address. Patients not being able to take part in a questionnaire study due to cognitive impairments were excluded from the study.

A priori, we calculated the necessary sample size to answer the primary research question. A sample size of 100 patients was adequate to specify limits of agreement within which 95% of paired differences of measurement approaches fall with an accuracy of 0.34 SD in the Bland-Altman plot.[24]

Measures

The SF-12, based on different recall periods, was used to assess patients' HRQoL. This generic instrument allows for comparisons across disease groups and has been validated in its German version.[25 26] It contains 12 items, which can be summarised into eight domains. The SF-12 standard version refers to the past 4 weeks (*SF-12 standard*); the acute version refers to the past week (*SF-12 acute*); and the daily version refers to the past 24 hours (*SF-12 daily*).

For use in economic evaluations, a preference-based utility index, the SF-6D, can be estimated based on seven SF-12 domains: physical functioning, role limitation (combined physical and emotional), bodily pain, vitality, social functioning, emotional role limitation and mental health. The preference-based algorithm uses health state valuations of the UK general population. The utility index ranges from 0 (worst health state) to 1 (best health state) and can be used for cost-effectiveness studies and for calculating QALYs (for further information, see online supplementary S1).[20]

In addition, we asked for the following sociodemographic characteristics: year of birth, gender, marital status, educational level, professional and housing situation, diagnosis, year of diagnosis and comorbidities.

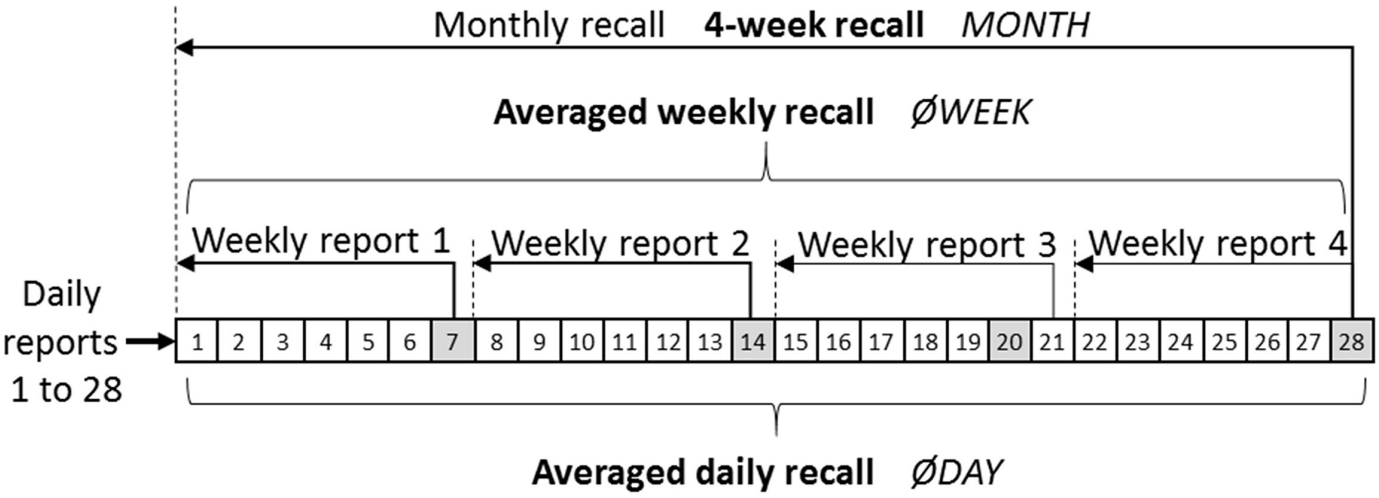

Figure 1 Data collection process for each patient.

Data collection

Two scientists recruited patients in the outpatient clinics between November 2017 and May 2018. Eligible patients were informed about the study and provided written informed consent. Subsequently, they completed a paper-based survey on sociodemographic characteristics and the online version of the SF-12 on a daily basis for 28 consecutive days: daily, considering the past 24 hours; weekly, considering the past 7 days; and at the last day of data collection, considering the past 4 weeks. This means that patients subsequently completed three versions of the SF-12 at the last day of data collection, each referring to a different recall period (*SF-12 daily*, *SF-12 acute* and *SF-12 standard*) (figure 1). For this, they received a daily automated invitation email. The time of the mailing was approximately 2 hours before the patient's individual bedtime to ensure a HRQoL assessment of the entire day. An additional text message reminder was offered on a voluntary basis. We asked patients to fill in the survey timely after receiving the invitation email but also permitted late completion until noon of the following day. If patients missed the last survey, including the 4-week recall survey, we reminded them about completion by telephone or email and allowed late completion. Patients received an expense allowance of up to €80 depending on the number of completed surveys. To control for day-of-week effect, the weekday of the start of data collection was assigned at random.

Data analysis

Data of patients who completed at least 14 of the 28 surveys, including the last one, were analysed. The preference-based SF-6D index was computed based on the 4-week recall (*MONTH*).[20] Missing values of single items (0.2%) were imputed by the weighed population mean of the total sample.[27] In addition, SF-6D indices for the daily and weekly HRQoL reports were computed and summary scores were calculated for each patient: (1) *ØDAY*, the mean of all SF-6D indices referring to the HRQoL of the past 24 hours, and (2) *ØWEEK*, the mean of all SF-6D indices referring to the HRQoL of the past week. This procedure resulted in three utility estimates for each individual patient (*ØDAY*, *ØWEEK* and *MONTH*), all relating to the same 28-day period.

Surveys that were completed later than noon of the following day were coded as missing; double entries were excluded. Sensitivity analyses were performed to detect the possible impact of late completion and missing surveys on the primary research question (agreement of *ØDAY* and *MONTH*).

To answer the primary research question, the agreement of *MONTH* and *ØDAY* was determined using the two-way mixed intraclass correlation coefficient (ICC) for single measures. We further analysed the agreement on the individual patient level by generating Bland-Altman plots.[24] These plots display statistical limits of agreement using the mean and the SD of the differences between two estimates, in this case, the difference between *MONTH* and *ØDAY* on the y-axis and the average of both estimates on the x-axis. In additional analyses, we determined the agreement between *MONTH* and *ØWEEK* and between *ØWEEK* and *ØDAY* using the same methods as described earlier.

Moreover, differences between *MONTH* and *ØDAY* were investigated using a paired sample t-test. Differences between both estimates were interpreted as constraints in recalling past states in retrospective assessments, that is, recall bias.

In order to explain recall bias (here only for the difference between *MONTH* and *ØDAY*), its association with different factors was investigated using Pearson correlation coefficients. First, to explore the extent to which patients were disproportionately influenced by the worst and the very last HRQoL report (peak-end effect[28]), the respective deviations from *ØDAY* were analysed for association with recall bias. Second, the association of recall bias with patient characteristics, that is, age, gender, educational level, working status, living situation, diagnosis, year of diagnosis, comorbidities and self-reported

HRQoL (*ØDAY*), was investigated. Last, patterns of dynamics in daily HRQoL reports were analysed for their association with recall bias. These patterns refer to the fluctuation of HRQoL over time; recall bias may vary depending on the degree of fluctuation. Three indicators of fluctuation that have previously been described by Houben and colleagues (2015) have been used in this study: (1) variability, (2) instability and (3) inertia.[29]

1. Variability describes the amplitude of patients' daily changes in HRQoL states. It is expressed as the within-person SD.
2. Instability characterises the magnitude of HRQoL shifts from 1 day to another. To quantify instability, differences between consecutive daily reports are squared and added up to the mean square successive difference.
3. Inertia indicates the extent to which HRQoL of 1 day can be predicted by the HRQoL of the previous day. This is expressed as the autocorrelation of daily values.

Finally, we performed a linear regression analysis to evaluate the combined predictive value of the factors described previously. A stepwise backward approach with probability to enter p=0.05 and probability to remove p=0.10 was chosen. As a sensitivity analysis, we also performed a regression model including all predictors.

The online survey tool QuestBack (Unipark, Cologne) was used to collect the data. Analyses were conducted using IBM SPSS Statistics V.23.

Patient and public involvement

The research question of the current observational study emerged because patients reported difficulties in recalling their HRQoL of a period in the past during a medical consultation or when participating in a research project. Our aim was to determine and quantify these difficulties. Patients or the public were not involved in the study design. Involvement of the public took place in the pretest phase of the online survey. A convenience sample of five healthy individuals judged the feasibility of the data collection process in general and the online survey in particular. According to the suggestions of healthy individuals, we decided to send daily invitation emails for completing the online survey at individualised times to account for individual preferences. For the same reason, we also decided to offer additional text message reminders. Finally, we offered the dissemination of individual study results to all patients who participated in the study.

RESULTS

Patient characteristics

To reach the predefined sample size of 100 participants, 124 potentially eligible patients were recruited. Twenty-two (17.7%) refused to participate; two patients (1.6%) completed less than 14 surveys (figure 2).

The final sample consisted of 50 patients with MS and 50 patients with psoriasis. The mean age of the total sample was 40.3 years (±11.95), and 63% were female. Of the 50 patients with MS, 8 were male and 42 were female. The psoriasis subgroup consisted of 29 men and 21 women. Descriptively, patients with MS tended to have a higher educational level and diagnosis was made more recently. Apart from that, subgroups were relatively similar (table 1).

Fifty-six patients completed all 28 surveys; 20 missed one survey only. The amount of missing surveys for the remaining 24 patients ranged between 2 and 12. Overall, the average number of missing surveys per case was 1.2 (±1.2). Of all 2681 completed surveys, 88.1% (n=2363) were completed in the evening of the respective day and

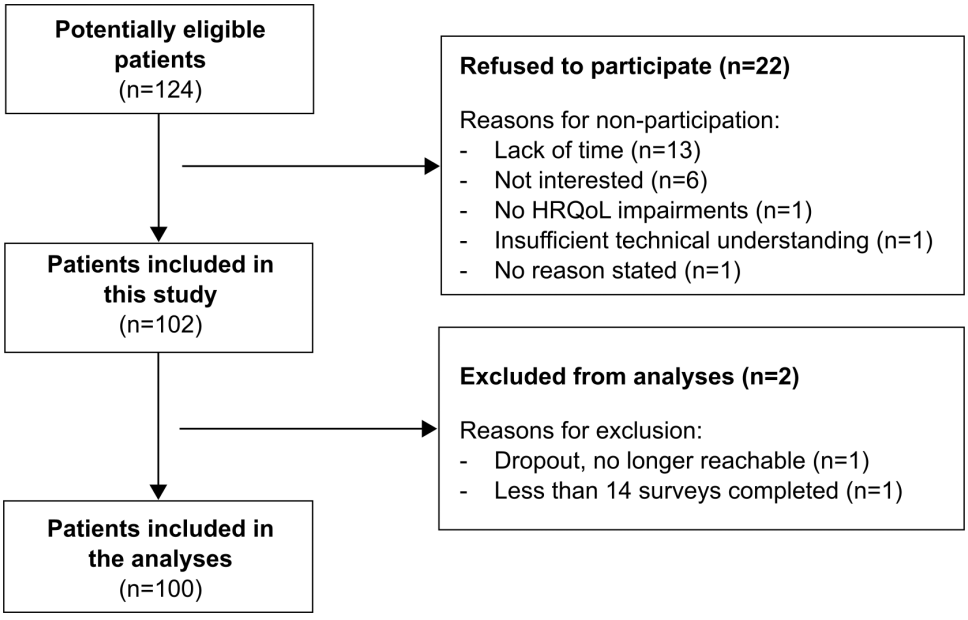

Figure 2 Flow diagram of the study participants. HRQoL, health-related quality of life.

Table 1 Demographic characteristics of the study participants

		Patients with MS (n=50)	Patients with psoriasis (n=50)	Total (N=100)
Gender, n (%)	Female	42 (84)	21 (42)	63 (63)
	Male	8 (16)	29 (58)	37 (37)
Age (years)	Mean±SD	37.2±10.25	43.4±12.82	40.3±11.95
	Median (range)	35 (20–62)	42.5 (21–67)	39 (20–67)
Educational level, n (%)	Low or medium	17 (34)	24 (48)	41 (41)
	High	33 (66)	26 (52)	59 (59)
Marital status, n (%)	Single	20 (40)	23 (46)	43 (43)
	Married/in a relationship	30 (60)	27 (54)	57 (57)
Working status (multiple responses possible), n (%)	Employed	35 (70)	35 (70)	70 (70)
	In training	6 (12)	8 (16)	14 (14)
	At home/unemployed	8 (16)	4 (8)	12 (12)
	Retired	10 (20)	7 (14)	17 (17)
Living situation, n (%)	Alone	8 (16)	11 (22)	19 (19)
	With family/friends/ partner	42 (84)	39 (78)	81 (81)
Time since diagnosis (years)	Mean±SD	8.6±7.56	17.9±13.74	13.3±11.99
	Median, range	7.5, 0–30	14, 1–65	10, 0–65
Comorbidities, n (%)	Yes	22 (44)	30 (60)	52 (52)
	No	28 (56)	20 (40)	48 (48)
SF-6D indices differing in recall period	*MONTH*, mean±SD	0.70±0.13	0.70±0.13	0.70±0.14
	ØWEEK, mean±SD	0.70±0.12	0.70±0.13	0.70±0.13
	ØDAY, mean±SD	0.74±0.12	0.73±0.14	0.74±0.13

ØDAY refers to the mean of SF-6D indices referring to the HRQoL of the past 24 hours; *ØWEEK* refers to the mean of SF-6D indices referring to the HRQoL of the past week; *MONTH* refers to the SF-6D index referring to the HRQoL of the past 4 weeks.
HRQoL, health-related quality of life; MS, multiple sclerosis; SF-6D, Short-Form Six-Dimension.

11.9% (n=318) were completed between midnight and noon of the following day. Sensitivity analyses indicated that exclusion of surveys with late completion and exclusion of patients with missing surveys did not change the results, considering the main research question.

Recall bias

The summary score of daily SF-6D was significantly (p<0.001) higher (*ØDAY*: 0.74±0.13) than the retrospectively rated SF-6D (*MONTH*: 0.70±0.14) with higher utility indices indicating better HRQoL. While differences between *ØDAY* and *ØWEEK* also reached statistical significance, differences between *ØWEEK* and *MONTH* did not. Absolute differences between indices, not taking into account the deviations' direction, were larger than the mean deviations. As expected, agreement between the three measurement approaches was high with the ICC ranging from 0.87 to 0.93 (table 2). In the sensitivity analyses, we also computed non-parametric correlations (Spearman's rho) and found similar results.

Bland-Altman plots display differences between the three measurement approaches on the individual patient level (figure 3). While for most patients (n=66) the retrospective judgement was more negative than the summary score of repeatedly daily reports (*ØDAY–MONTH*>0), there were also 30 patients for whom the opposite could be observed. The even distribution of differences along the x-axis indicates that differences between measures did not depend on the health state itself; that is, a negative or a positive mean SF-6D was not associated with greater recall bias. Overall, the range of differences was greatest between *ØDAY* and *MONTH* and smallest between *ØDAY* and *ØWEEK*.

Factors affecting recall

Recall bias, measured by the absolute difference between *MONTH* and *ØDAY,* decreased with age (r=−0.24, p=0.02) and increased with higher self-reported HRQoL (*ØDAY*: r=0.17, p=0.03). Only self-reported HRQoL remained a significant predictor in the stepwise backwards regression model. Correlations with the remaining patient characteristics such as the underlying disease (MS vs psoriasis), gender or educational level were non-significant. Recall bias was also associated with the extremity of the 'peak', that is, the deviation of the worst daily HRQoL report from the summary score *ØDAY* (r=0.52, p<0.001),

Table 2 Agreement between measurement approaches differing in recall period (N=100)

SF-6D indices differing in recall period	Paired sample t-test				ICC single measure	
	Absolute difference	Mean difference (95% CI)	P value	d	ICC (95% CI)	P value
ØDAY–MONTH	0.05	0.04 (0.02 to 0.05)	<0.001	0.55	0.87 (0.81 to 0.91)	<0.001
ØDAY–ØWEEK	0.04	0.03 (0.02 to 0.04)	<0.001	0.70	0.93 (0.90 to 0.95)	<0.001
ØWEEK–MONTH	0.03	0.01 (0.00 to 0.02)	0.38	0.09	0.92 (0.89 to 0.95)	<0.001

ØDAY refers to the mean of SF-6D indices referring to the HRQoL of the past 24 hours; *ØWEEK* refers to the mean of SF-6D indices referring to the HRQoL of the past week; *MONTH* refers to the SF-6D index referring to the HRQoL of the past 4 weeks. Absolute difference between SF-6D indices disregard the direction of the deviation.
d, effect size parameter Cohen's d for paired sample t-test; HRQoL, health-related quality of life; ICC, intraclass correlation coefficient;SF-6D, Short-Form Six-Dimension.

and with two measures of patterns of dynamics, namely, variability (0.60, p<0.001) and instability (0.65, p<0.001). Thus, recall bias is more likely if patients experience high fluctuation of HRQoL over time (table 3).

Results of the regression analyses further underpinned the impact of fluctuation of HRQoL over the recall period. Variability and instability in the stepwise model and instability in the full model were influencing predictors of absolute *ØDAY–MONTH* difference in the regression models. Non-employment (in both models) and higher self-reported HRQoL (*ØDAY*) (in the stepwise model) were further significant predictors. Overall, the predictors explained 47% (stepwise) and 43% (full model) of variance regarding the *ØDAY–MONTH* difference (p<0.001, table 3).

DISCUSSION

The aim of this study was to assess the agreement between preference-based HRQoL reports with different recall periods. The main finding was that in patients with psoriasis or MS, retrospective reports of the past 4 weeks were not identical to the average of repeated daily reports. Recall bias seemed to be present in the SF-6D answers. On the group level, the retrospective reports were slightly more negative than the average of daily reports. This suggests that patients with MS or psoriasis tend to give more weight to negative experiences in the past or to remember negative emotions better. On the individual level, we observed deviations in both directions, with retrospective underestimation being more prevalent than overestimation. Also, deviation was greater in patients experiencing higher variability of HRQoL over time.

Recall bias on the group level

The mean difference between the repeated daily HRQoL reports (*ØDAY*) and the retrospective reports of the past 4 weeks (*MONTH*) had a magnitude similar to the minimally important difference[30] identified for the SF-6D in numerous study populations.[31 32] Thus, mean differences between *ØDAY* and *MONTH* were small, but the effect size was medium and differences could be clinically meaningful. A similar difference between *ØDAY* and *ØWEEK* based on medium effect size reveals that recall bias should already be considered for recall periods of 1 week. Hence, economic evaluations based on the SF-6D both in its standard (4-week recall) and in its acute (1-week recall) versions could be slightly impacted by recall bias.

In this study, recall bias may even be underestimated, as the study design may have enhanced memory and thereby diminished recall bias. Patients completed surveys on a daily basis. Thereby, they intensively focused on evaluating their own HRQoL during data collection, which might have facilitated recollection. Recall bias may therefore be greater when data are collected retrospectively only, as commonly done in research, economic evaluations and clinical practice. In addition, recollection could be worse in respondents who do not have a chronic disease. Treatment of chronic diseases usually pursues HRQoL improvement as an important treatment goal; therefore, patients with a chronic disease may think about their HRQoL more often than healthy individuals or patients with acute diseases. This may improve recollection.

Due to the subjective nature of HRQoL, statements on the accuracy of retrospective reports remain challenging. There is no gold standard and thus no true value to compare HRQoL data to, but it is highly probable that memory influences data accuracy.[33] The present study supports theories that memories on past experiences decline over time, fostering recall bias in the retrospective measurement of subjective constructs.[34–38] Consequently, diary data are assumed to be less affected by recall bias and are therefore commonly used for the validation of retrospective patient-reported outcomes.[39] Findings of such validations—in line with the results of the present study—suggest a general overestimation of negative experiences for patient-reported outcomes such as pain or well-being.[16 40–42]

Recall bias on the individual level

While group-level results suggested small mean differences and an agreement of measurement approaches sufficient for research purposes, discrepancies on the individual level were greater and bidirectional: some patients markedly underestimated retrospective HRQoL,

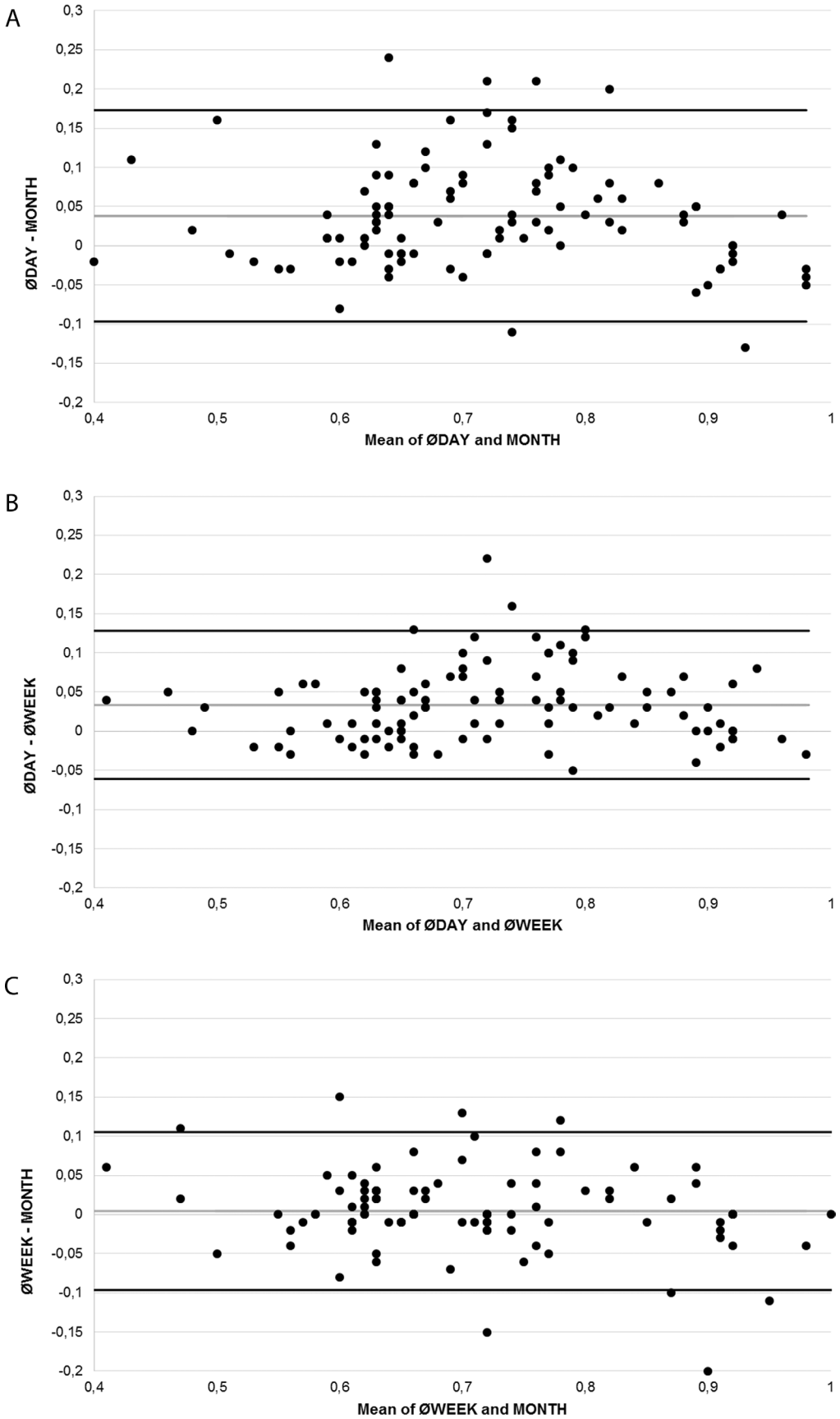

Figure 3 Bland-Altman plots for SF-6D indices differing in recall period. Legend: Bland-Altman plots for comparison between recall periods: (A) *ØDAY* and *MONTH*, (B) *ØDAY* and *ØWEEK*, and (C) *ØWEEK* and *MONTH*. The x-axis displays the mean of two indices; the y-axis displays the difference between them. The mean difference (grey line) and the 95% limits of agreement calculated by 1.96 SD (black lines) are marked. SF-6D, Short-Form Six-Dimension.

Table 3 Associations of patient characteristics, peak-end effect and patterns of dynamics with absolute difference between *MONTH* and *ØDAY* (N=100)

	Bivariate correlation		Linear regressions: Stepwise backwards method		Linear regressions: Enter method	
	R	P value	Beta	P value	Beta	P value
Patient characteristics						
Age (years)	**–0.24**	**0.02**	–	–	–0.10	0.37
Gender (ref: female)	–0.02	0.88	–	–	–0.05	0.59
Diagnosis (ref: multiple sclerosis)	–0.06	0.60	–	–	0.01	0.93
Educational level (ref: low)	0.17	0.09	–	–	0.11	0.25
Working status (ref: not employed)	–0.08	0.44	**–0.22**	**0.01**	**0.20**	**0.02**
Living situation (ref: alone)	0.10	0.34	–	–	–0.04	0.63
Time since diagnosis (years)	0.01	0.95	–	–	0.10	0.33
Comorbidities (ref: no)	–0.06	0.29	–	–	–0.01	0.94
SF-6D index, ØDAY (mean)	**0.20**	**0.003**	**0.17**	**0.03**	0.13	0.21
Peak-end effect						
Worst daily report (deviation from mean)	**0.52**	**<0.001**	–	–	0.11	0.46
Last daily report (deviation from mean)	0.17	0.09	–	–	–0.03	0.77
Patterns of dynamics						
Variability (within-person SD)	**0.60**	**<0.001**	**0.27**	**0.05**	0.24	0.26
Instability (MSSD)	**0.65**	**<0.001**	**0.43**	**<0.001**	**0.37**	**0.02**
Inertia (autocorrelation)	–0.14	0.18	–	–	–0.07	0.42
R^2	–		**0.49**		**0.52**	
Adjusted R^2	–		**0.47**		**0.43**	
SE	–		**0.04**		**0.04**	

ØDAY refers to the mean of SF-6D indices referring to the health-related quality of life of the past 24 hours.
Significant values in bold
MSSD, mean square successive difference; ref, reference category; SF-6D, Short-Form Six-Dimension.

others overestimated it. This is why the absolute deviation of *ØDAY* and *MONTH* was larger than the mean deviation. Thus, recall bias is of greater importance with regard to individual patient reports. In clinical practice, individual HRQoL reports are used to comprehend the patients' experiences and to include them in the decision-making process.[2] For individual consultations, short recall periods may therefore be more suitable for gaining a less distorted impression on the patient's impairments in HRQoL.[34]

A differentiated view on particular subgroups of patients

Recall bias was more likely to occur in particular subgroups of the study population. Patients who experienced considerable changes in HRQoL over time tended towards larger recall bias. This indicates that single daily reports are not valuated equally in retrospective assessments. The phenomenon of valuing experiences disproportionally has also been observed for self-reports on other subjective constructs. In particular, retrospective patient-reported outcomes seem to be disproportionally influenced by the worst and the very last experience.[16 37] In our study, we could confirm the impact of the worst state of HRQoL on the agreement but not the impact of the very last day.

Furthermore, we found that diagnosis and gender were not associated with recall bias, whereas employment status was: employed patients were less likely to experience recall bias. A reason could be that a regulated daily routine facilitates memories on past experiences. Overall, interindividual variance in recall bias could be explained to a large extent by indicators of dynamics and employment status. Overall, however, subgroup analyses must be interpreted with caution. Bivariate correlation analyses and linear regression analyses indicate a tendency only and need to be confirmed in further analyses.

Strengths and limitations

Our findings should be viewed in the context of some strengths and limitations. Recall bias was analysed in patients with two specific chronic conditions and for a single utility measure only, which limits generalisability. It should also be noted that our study population was not selected to be representative to all patients with psoriasis and MS. This could be the reason why health states of both patient groups were evaluated similarly in our study, while disability weights in the Global Burden of Disease Study were greater for patients with MS than for patients with psoriasis.[21] In addition, both groups were similar in terms of numerous sociodemographic characteristics and differed mainly in terms of sex ratio and time since diagnosis.

In this study, we analysed recall bias with respect to the SF-6D and focused on the total utility index only. We did not distinguish between different domains of HRQoL and therefore cannot make any statements about whether recall bias is larger for some domains than for others.

In general, although data were relatively complete (ie, few missing surveys and few missing values within single surveys), some surveys were missing due to problems with delivery of single invitation emails and the survey software. Due to software configuration problems, patients could skip single answers within a single survey, although we intended to include mandatory items only. Apart from these rather minor technical problems, electronic data collection was a major strength of our study. Contrary to traditional paper-based diaries, the electronic data collection enabled monitoring of incoming surveys and prevented retrospective completion of diary entries.[43]

Practical implication

We found that recall bias impacts retrospective utility estimates. On the group level, however, bias was relatively small. Thus, for research purposes and in particular for economic evaluations, where the group level is of major interest, a 4-week recall period could be considered appropriate. In this context, it needs to be considered that, for particular groups, specifically for patients who are expected to experience high fluctuation of HRQoL over time or for patients with no regular daily routine, recall bias could be of greater significance. For those groups, data collection based on diaries may be more appropriate. Using diaries could also be an opportunity to combat recall bias in clinical practice, where the individual patient is the focus of consideration. However, extra burden on patients of completing a survey daily instead of once for a retrospective time period should not be underestimated.[8]

CONCLUSIONS

Recall bias should not be disregarded in retrospective HRQoL assessments. While bias was relatively small on the group level, it was more severe on the individual level. Therefore, it is essential to distinguish between purposes of data collection. When using summary scores of a population to determine treatment utility in economic evaluations, retrospective overestimation and underestimation of single patients almost cancel out one another. Caution is advised with interpretation of single utility scores or HRQoL reports that are used as a basis for treatment decisions in clinical practice.

Acknowledgements We thank all patients who participated in the study. Additionally, we are very grateful to the clinicians at the outpatient clinics for multiple sclerosis and psoriasis of the University Medical Center Hamburg-Eppendorf (UKE) for their support in the recruitment of participants.

Contributors All authors substantially contributed to the conception and design of the study and the interpretation of the data. JT and VA were responsible for data acquisition, JT, VA and CB were involved in the data analysis. JT drafted the work; VA, CH, MA and CB commented on it and revised it critically. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Funding This work was supported by the Federal Ministry of Education and Research of Germany grant number 01EH160 1B HCHE.

Competing interests None declared.

Patient consent for publication Obtained.

Ethics approval The study was carried out in accordance with the code of ethics of the Declaration of Helsinki and was approved by the ethics committee of the Medical Association Hamburg (reference number PV5508). Each participant provided a written informed consent before participation in the study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Janine Topp http://orcid.org/0000-0002-7105-3244

REFERENCES

1 Mokkink LB, Terwee CB, Gibbons E, *et al*. Inter-rater agreement and reliability of the COSMIN (consensus-based standards for the selection of health status measurement instruments) checklist. *BMC Med Res Methodol* 2010;10:82.
2 Refolo P, Minacori R, Mele V, *et al*. Patient-reported outcomes (PROs): the significance of using humanistic measures in clinical trial and clinical practice. *Eur Rev Med Pharmacol Sci* 2012;16:1319–23.
3 Neumann PJ, Goldie SJ, Weinstein MC. Preference-based measures in economic evaluation in health care. *Annu Rev Public Health* 2000;21:587–611.
4 Fayers PM, Machin D. *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. 3rd edn. Chichester: John Wiley & Sons, 2007.
5 Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ* 1986;5:1–30.
6 Drummond MF, Sculpher MJ, Claxton K, *et al*. *Methods for the economic evaluation of health care programmes*. 4th edn. United Kingdom: Oxford University Press, 2015.
7 European Network for Health Technology Assessment (EUnetHTA). Methods for health economic evaluations - a guideline based on current practices in Europe 2015.
8 Stull DE, Leidy NK, Parasuraman B, *et al*. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. *Curr Med Res Opin* 2009;25:929–42.
9 Keller SD, Bayliss MS, Ware JE, *et al*. Comparison of responses to SF-36 health survey questions with one-week and four-week recall periods. *Health Serv Res* 1997;32:367–84.
10 Reeve BB, Wyrwich KW, Wu AW, *et al*. ISOQOL recommends minimum standards for patient-reported outcome measures used in

patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;22:1889–905.
11 Norquist JM, Girman C, Fehnel S, *et al*. Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration. *Qual Life Res* 2012;21:1013–20.
12 Coughlin SS. Recall bias in epidemiologic studies. *J Clin Epidemiol* 1990;43:87–91.
13 Salovey P, Sieber WJ, Jobe JB, *et al*. The Recall of Physical Pain. In: Schwarz N, Sudman S, eds. *Autobiographical memory and the validity of retrospective reports*. New York: NY: Springer, 1994.
14 Broderick JE, Schwartz JE, Vikingstad G, *et al*. The accuracy of pain and fatigue items across different reporting periods. *Pain* 2008;139:146–57.
15 Erskine A, Morley S, Pearce S. Memory for pain: a review. *Pain* 1990;41:255–65.
16 Stone AA, Schwartz JE, Broderick JE, *et al*. 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.
17 Kahneman D, Fredrickson BL, Schreiber CA, *et al*. When more pain is preferred to less: adding a better end. *Psychol Sci* 1993;4:401–5.
18 Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Qual Life Res* 2003;12:599–607.
19 Clarke PM, Fiebig DG, Gerdtham U-G. Optimal recall length in survey design. *J Health Econ* 2008;27:1275–84.
20 Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care* 2004;42:851–9.
21 Global Burden of Disease Collaborative Network. *Global burden of disease study 2016 (GBD 2016) disability weights*. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2017.
22 Strober B, Greenberg JD, Karki C, *et al*. Impact of psoriasis severity on patient-reported clinical symptoms, health-related quality of life and work productivity among US patients: real-world data from the Corrona psoriasis registry. *BMJ Open* 2019;9:e027535.
23 von Elm E, Altman DG, Egger M, *et al*. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* 2007;4:e296.
24 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
25 Ware J, Kosinski M, Keller SD. A 12-Item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Med Care* 1996;34:220–33.
26 Gandek B, Ware JE, Aaronson NK, *et al*. Cross-Validation of item selection and scoring for the SF-12 health survey in nine countries: results from the IQOLA project. International quality of life assessment. *J Clin Epidemiol* 1998;51:1171–8.
27 Perneger TV, Burnand B. A simple imputation algorithm reduced missing data in SF-12 health surveys. *J Clin Epidemiol* 2005;58:142–9.
28 Fredrickson BL. Extracting meaning from past affective experiences: the importance of peaks, ends, and specific emotions. *Cogn Emot* 2000;14:577–606.
29 Houben M, Van Den Noortgate W, Kuppens P. The relation between short-term emotion dynamics and psychological well-being: a meta-analysis. *Psychol Bull* 2015;141:901–30.
30 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407–15.
31 Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual Life Outcomes* 2003;1:4.
32 Kaplan RM. The minimally clinically important difference in generic utility-based measures. *COPD* 2005;2:91–7.
33 Gutek BA. On the accuracy of retrospective attitudinal data. *Public Opin Q* 1978;42:390–401.
34 Schneider S, Stone AA. Ambulatory and diary methods can facilitate the measurement of patient-reported outcomes. *Qual Life Res* 2016;25:497–506.
35 Conway MA, Pleydell-Pearce CW. The construction of autobiographical memories in the self-memory system. *Psychol Rev* 2000;107:261–88.
36 Rubin DC, Wenzel AE. One hundred years of forgetting: a quantitative description of retention. *Psychol Rev* 1996;103:734–60.
37 Robinson MD, Clore GL. Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychol Bull* 2002;128:934–60.
38 Lovalekar M, Abt JP, Sell TC, *et al*. Accuracy of recall of musculoskeletal injuries in elite military personnel: a cross-sectional study. *BMJ Open* 2017;7:e017434.
39 Stone AA, Shiffman S, Atienza A. *The science of real-time data capture: Self-Reports in health research*. New York: Oxford University Press, 2007.
40 Rydén A, Leavy OC, Halling K, *et al*. 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;19:829–33.
41 Schneider S, Broderick JE, Junghaenel DU, *et al*. Temporal trends in symptom experience predict the accuracy of recall pros. *J Psychosom Res* 2013;75:160–6.
42 Bennett AV, Amtmann D, Diehr P, *et al*. Comparison of 7-day recall and daily diary reports of COPD symptoms and impacts. *Value Health* 2012;15:466–74.
43 Stone AA, Shiffman S, Schwartz JE, *et al*. Patient non-compliance with paper diaries. *BMJ* 2002;324:1193–4.