Quality of Life Research (2022) 31:3433–3445
https://doi.org/10.1007/s11136-022-03198-9

How do the number of missing daily diary days impact the psychometric properties and meaningful change thresholds arising from a weekly average summary score?

Pip Griffiths[1] · **Abi Williams**[2] · **Elaine Brohan**[3]

Accepted: 5 July 2022 / Published online: 5 August 2022

Abstract

Purpose Quality of life research often collects daily information and averages this over a week, producing a summary score. When data are missing, arbitrary rules (such as requiring at least 4/7 observations) are used to determine whether a patient's summary score is created or set to missing. This simulation work aimed to assess the impact of missing data on the estimates produced by summary scores, the psychometric properties of the resulting summary score estimates and the impact on interpretation thresholds.

Methods Complete longitudinal data were simulated for 1000 samples of 400 patients with different day-to-day variability. Data were deleted from these samples in line with missingness mechanisms to create scenarios with up to six days of missing data. Summary scores were created for complete and missing data scenarios. Summary score estimates, psychometric properties and meaningful change estimates were assessed for missing data scenarios compared to complete data.

Results In most cases, the 4/7 day rule was supported, but this depended on daily variability. Fewer days of data were sometimes acceptable, but this was also dependent on the proportion of patients with missing data. Tables and figures allow researchers to assess the potential impact of missing data in their own studies.

Conclusions This work suggests that the missing data rule used to create summary scores impacts on the estimate, measurement properties and interpretation thresholds. Although a general rule of 4/7 days is supported, the way the summary score is derived does not have a uniform impact across psychometric analyses. Recommendations are to use the 4/7 rule, but plan for sensitivity analyses with other missing data rules.

Keywords Daily diary data · Missing data · Summary score · Psychometric properties · 4/7 days

Introduction

When a health condition can vary rapidly in a short space of time, assessment at periodic clinic visits (such as every 6 months) may not be the most appropriate way of assessing change. This is a feature of many aspects of how a patient feels or functions, for example, in the assessment of mood, respiratory symptoms, pruritus and pain. A patient-reported outcome (PRO) measure with long recall (e.g. 7 days or more) which covers a period where a health state has fluctuated may lead to an inaccurate representation of the construct being measured. Indeed, when comparing daily diary assessments to assessments completed at clinic visits in asthma, daily measurement has been shown to be more sensitive to detecting changes in patient health over time [1]. Additionally, retrospective reports of pain [2] and COPD [3] have been shown to be more likely to be inflated (i.e. worse symptom report) compared to daily report.

Daily measurement can now be conducted using electronic diaries, which can increase both the integrity and accuracy of the data collected [4] compared with paper completion. For example, prior to the invent of electronic data collection, where each entry is automatically time-stamped for assurance of integrity [4], research showed that participants tended to complete the diary questionnaire retrospectively by filling out

✉ Pip Griffiths
pip.griffiths@iqvia.com

[1] Patient Centric Services, IQVIA France, Tour D2, 17 bis Passerelle des Reflets, 92400 Courbevoie, France

[2] Patient Centric Services, IQVIA UK, 3 Forbury Place, 23 Forbury Rd, Reading RG1 3JH, UK

[3] Formerly of Patient Centered Outcomes, Adelphi Values, Bollington, UK

many daily assessments at one timepoint [5], a phenomenon known as the "parking lot effect." This leads to what appears to be a high completion rate, but it is based on retrospective rather than daily completion. This may result in potentially inaccurate data due to recall effects. Electronic data collection can, on the other hand, prevent the completion of daily diaries outside of an allotted time window, set by the investigator. This allows for data integrity to be preserved but can lead to missing data when participants do not, for some reason, complete the diary in the allotted time [6].

Although good electronic daily diary compliance rates (> 90%) have typically been reported [5, 7], accounting for missing data is still an important consideration with this mode of administration. Research has shown that PRO measure completion can vary, and specifically decline, over the course of a trial. This is potentially due to increased technical or condition-related issues completing the diary the longer patients are enrolled [8]. High completion should therefore not be taken for granted. Missing data in daily diary studies can occur for a broad range of reasons including forgetting, being too busy, technical malfunction of the device, being too unwell to complete the diary [9] or simply not wanting to complete the questionnaire each day. The level of burden placed on participants is also an important consideration, particularly for data that are event-driven rather than daily and may require multiple completions over a day, e.g. seizure data. Due to the variability inherent in event-driven diary data, this paper focuses on daily diary data though many of the considerations will also apply to event-driven diary data, where participants are required to complete the diary when a certain condition-related event occurs (for example, when suffering an episode of dysphagia). The reasons for missing daily diary assessment all link to statistical definitions of missing data that are often considered when analysing PRO results which contain missing records (Table 1).

Although these missing data mechanisms have been well documented in the literature, missing data still pose a challenge for statistical analysis, see [10]. These challenges mainly arise due to the difficulty in assigning missingness in a dataset to one of the categories presented in Table 1. The true underlying value of any unobserved missing data is by definition unknown, meaning it can be difficult to distinguish between missing data mechanisms (specifically MAR and MNAR) analytically. To account for this, it is recommended that multiple analysis approaches are used, each with different missing data assumptions, in order to assess the robustness of any results to the assumptions made[11]. Although these recommendations have been made for clinical trial analyses, similar considerations could, and perhaps should, also apply for psychometric analyses.

Missing data can lead to a reduction in the precision of a treatment effect and in the most severe cases can lead to biased estimates, especially if the reasons for the missing diary assessments are believed to be linked to the condition being assessed [12]. Box 1 shows an illustrative example of this, and the problems that arise from missing data. Missingness is of particular concern when assessing data arising from daily diary collection; this is due to the way the data are summarized in a trial setting. For example, to account for the daily variability in the condition of interest, daily assessment is often summarized over a time period such as a week by taking the mean of all assessments in that week [13]. This is a particularly typical way to assess diary data in clinical trial work where it is preferable to have a summary score which is linked to a participant's clinic visit. This summary score is used to measure change over time to assess treatment efficacy. For daily diary data, a common practice rule of thumb for dealing with missing data is to average over the remaining data in the time period. There is sometimes a rule associated with this, such as only using subjects who provide at least the majority of diary days in that time period (e.g. at least 4 days out of 7). When this kind of rule is applied, participants with fewer days of data than this cut off do not have an average calculated from their available data and instead are awarded a missing value for the summary score. This means that these participants are often not included in subsequent analyses.

Table 1 Examples of missingness mechanisms and their consequences

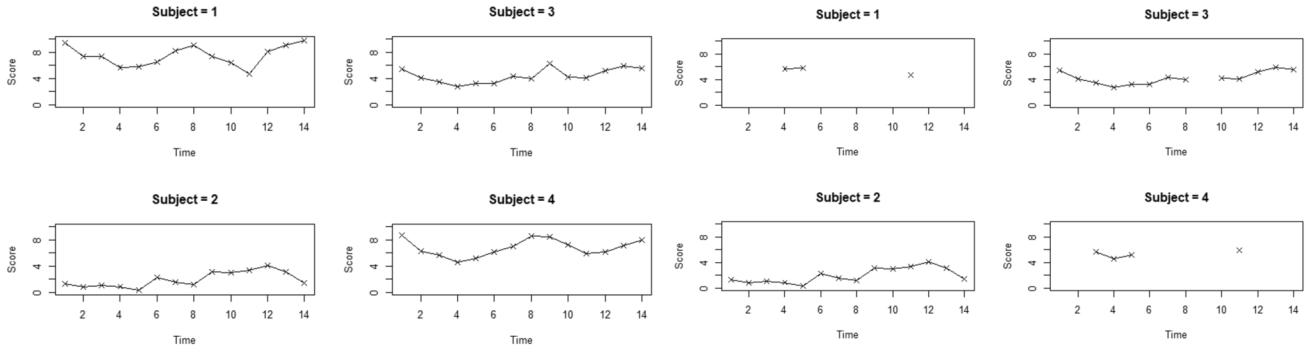
Missingness mechanism	Definition	Daily diary example (influences missingness)	Issues
MCAR (Missing Completely at Random)	Missingness is unrelated to the construct being observed	Technical malfunction Forgetting[a]	Increased variability
MAR (Missing at Random)	After conditioning on observed data, missingness is not related to the unobserved missing value	Other daily/continuous measure (e.g. heart rate) Previous assessments Too busy[b]	Increased variability and potential bias[c]
MNAR (Missing Not at Random)	The missingness is directly related to the unobserved missing value	Too sick Too busy[b]	Increased variability and bias

[a]Provided the construct under investigation is not memory

[b]Being too busy could relate to the construct of interest if the patient health is improving and is therefore able, for example, to return to work

[c]Bias if observed data relating to the missingness are not included in the model

Box 1 content



As an extreme illustrative example, imagine four clinical trial patients reporting their daily pain. On the left, these four patients have complete 14 days of daily reported pain data on a 0–10 scale, with 0 being no pain and 10 being the worst pain imaginable. Here, we can see that pain increases on some days and subsides on others. On the right-hand side, we show a scenario for the same four patients' data where higher pain scores are Missing Not at Random (in this case, arbitrarily deleting a score of 6 or more on a 0–10 scale for this example). This can be conceived as a situation where the pain a patient is experiencing is so debilitating, it is preventing them from completing the daily diary and reporting their pain. As you can see, this leads to the most severe scores being omitted from the analysis dataset. This then has follow-on consequence for the patients' observed pain score and mean estimates at the individual and group level, with each showing *less* observed pain than the patient experienced. When missingness related to severe scores such as in this example exists, the most severe scores will not be observed and any mean estimates will be biased towards less severe scores.

Subject	14-Day Average	
	Complete Data	Missing Data Scenario
1	7.48	5.36
2	1.99	1.99
3	4.42	4.28
4	6.78	5.86
Group Mean	5.17	4.37

This pattern could be reversed, if patients were less likely to complete the diary on "good days" where they have less severity and can continue their life unabetted by their condition.

Until recently, there was little evidence for this rule of thumb and the choice of averaging participant data provided that more than half the days of data were present was mainly based on the researcher's prior experience. Initial simulation work in this area showed that group-level estimates which were based on a summary score created when patients had at least three days of daily diary data were closer to the true, simulated, score in terms of variability and bias than when the patients with missing data were excluded from the estimate [14]. In addition, reliability of the summary score over time was more likely to be preserved when summary scores included participants with as few as 3 days of diary data, rather than excluding participants who have missing data [15], Floden et al., in preparation.

Given that the rule of thumb currently employed to create summary scores from missing data over a week may not be the optimal strategy when it comes to the accuracy of the weekly summary score estimate or its reliability over time, it is important to understand whether other properties of such summary scores are better served by using all available data, or by excluding participants who do not have complete data. This is a particularly pertinent issue for clinical trial work where the reliability of the summary score, the validity of the summary score (i.e. how accurately it assesses the construct of interest [16]), the ability of the summary score to detect change and the interpretation of that score are crucial to properly assessing the efficacy of a potential new investigational medical product.

This work used a simulation approach to test these questions across different types of missing data and across scenarios which represented diseases with low, moderate, and high daily variability. The following hypotheses were specifically tested:

(1) How many days of daily diary data are needed to create a non-biased score estimate for different missingness mechanisms, different variabilities of daily scores, and different proportions of patients with missing data
(2) How are the measurement properties of a daily diary impacted by creating a summary score from records with missing data, specifically, reliability, validity, and ability to detect change.
(3) How the group-level and individual-level interpretation thresholds are impacted by creating a summary score from records with missing data.

Methods

Simulation procedures

Simulations were conducted in SAS version 9.4. The seed was set as 12345.

The simulation was based on three different populations as a motivating example. These three populations represented differences in the day-to-day variability of different diseases and included populations with low daily variability, moderate daily variability and high daily variability (Table 2). These correlation structures were based on the authors prior experience and theoretical assumptions. An example of a condition which may have low variability is chronic pain, moderate variability may exist in a dermatological condition, whereas a respiratory condition involving symptom flares may have high daily variability.

For each of the three variability conditions, 1000 samples each with 400 records (representing participants), split into 2 groups (representing a treatment group and a control group) were generated. Each participant had data simulated at two timepoints (baseline and follow-up) and each timepoint consisted of 7 days of data (observations). For baseline and follow-up for the control group, and for baseline for the treatment group, a multivariate normal distribution with a mean of 5 and a standard deviation (SD) of 2 was used to generate the data. As such, for the treatment group at the follow-up timepoint, a multivariate normal distribution was used with a mean of between 4.350 and 4.575 and a SD of 2. PROC SIMNORMAL in SAS was used to simulate the data to the above specifications, using positive definite versions of the Table 2 correlation matrices as the input dataset (see supplementary material for further description and presentation of the full positive definite matrices [17]).

To clarify, this process led to two timepoints of data for both the treatment and control groups. The low correlation of 0.30, specified between Day 7 of baseline and Day 1 of follow-up led to a plausible scenario where there was some weak correlation between the two timepoints. To verify the relationship between the two timepoints which were assumed to have some undetermined period of time between them, a correlation analysis was performed between

Table 2 Correlation structure for simulating daily variability

Time	Day	Low variability							Moderate variability							High variability						
Baseline		1*	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
	1																					
	2	.80							.70							.40						
	3	.70	.80						.55	.70						.25	.40					
	4	.60	.70	.80					.40	.55	.70					.10	.25	.40				
	5	.50	.60	.70	.80				.25	.40	.55	.70				.10	.10	.25	.40			
	6	.40	.50	.60	.70	.80			.10	.25	.40	.55	.70			.10	.10	.10	.25	.40		
	7	.30	.40	.50	.60	.70	.80		.10	.10	.25	.40	.55	.70		.10	.10	.10	.10	.25	.40	
Follow-up[ł]	1	.00	.00	.00	.00	.00	.00	.30	.00	.00	.00	.00	.00	.00	.30	.00	.00	.00	.00	.00	.00	.30
	2–7	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

*The horizontal column labels for days 1–7 represent Days 1–7 of baseline

ł The correlation (0.30) between baseline Day 7 and follow-up Day 1 was set so as to simulate 2 weeks of data which were separated by time. The correlation structure for follow-up Day 1–7 was the same as for baseline. This means that there were two "weeks" of data which independently had a similar correlation structure, but the 0.30 correlations between baseline Day 7 and follow-up Day 1 dictated the overall relationship between the two "weeks" and ensured some weak association over time

the mean of all daily scores at baseline and follow-up. These values had a low, but consistent relationship in line with the manner in which these two timepoints were simulated ($r = 0.1$–0.2).

These simulated scores were assumed to be the underlying "latent" score for each patient. Rounding these values allowed for systematic "error" to be added to the latent scores. Rounding led to values between − 1 and 13, which were then restrained to lie between 0 and 10 thus creating the observed scores on a 0–10 numeric rating scale (NRS). For each patient in each simulated dataset, a summary score estimate was created for baseline and follow-up as the mean of all 7 days of observed scores. This is referred to throughout as the Observed Summary Score.

Additional variables simulated as part of this analysis are detailed in supplementary material 2.

Introduction of missing data

Missing data were introduced into each scenario at follow-up and followed the three missingness mechanisms (missing completely at random [MCAR], missing at random [MAR], and missing not at random [MNAR]) [10, 12].

The different missingness mechanisms were used to decide both on the participants selected to have datapoints deleted and the datapoints within participants' records that would be deleted. Participants selected to receive missing data were based on:

(1) MCAR: Participants were selected completely at random
(2) MAR: Participants were selected based on their Observed Summary Score at baseline, with participants with a more severe response at baseline being more likely to be selected.
(3) MNAR: Participants were selected based on their Observed Summary Score at follow-up.

Datapoints selected to receive missing data were based on:

(1) MCAR: datapoints were selected completely at random
(2) MAR: Datapoints were weighted for selection based on the severity of the previous day's severity. The higher the value of the previous days observed score, the more likely the datapoint would be selected for deletion. When the previous day was missing, the next available day was used. When there was no previous day, the baseline Observed Summary Score was used as a weight.
(3) MNAR: datapoints were selected based on the severity of the observed datapoint for that day. The higher the value of that day's observed score, the more likely the datapoint would be selected for deletion.

This simulation setup is expected to lead to a negative bias in scores for MAR and MNAR. This is because the *more severe* responses are unobserved through missingness in this case. The opposite pattern maybe be expected if, for example, patients were improving and therefore less likely to complete their diary because they were feeling better and continuing with life. In this case, the bias would be expected to be positive. This could conceptually be the case in a COVID-19-related diary, where patients who are feeling better return to work and forget to complete their diary.

Missing data were introduced in increasing proportions. This means that either 10%, 20%, 30%, or 40% of the sample were missing either 1, 2, 3, 4, 5, 6, or all 7 days of data. When all 7 days of data were missing for a proportion of the sample, this is akin to an analysis which only uses participants with complete data. In total (not including the full dataset arising from the initial simulation), this led to 28 scenarios for each of the 3 daily variability conditions for each of the missingness mechanisms, for a total of 252 scenarios involving some level of missing data. This led to a total of 252,000 datasets, each with 400 patients observed scores.

For each participant in each scenario, the Observed Summary Score for follow-up was recalculated as the mean of all remaining datapoints. When all 7 days of data had been removed, no Observed Summary Score was created as there were no data from which to create a summary score. In this case, the patients Observed Summary Score was set to missing.

These metrics were used in the assessment of the performance and psychometric properties of the Observed Summary Score under different missingness conditions.

Assessment of observed summary score estimate, psychometric properties, and meaningful change analysis performance

Prior to showing the impact on psychometric properties, the performance of the Observed Summary Score created from records with missing data was assessed in terms of bias and root mean squared error. Estimate bias was assessed as the difference between the Observed Summary Score at follow-up created from partial datasets (i.e. datasets with missing data) and the "true" Observed Summary Score which was created from the complete data before data were made missing. This was calculated for each sample in each missingness scenario:

- Mean of (Participant follow-up Observed Summary Score from partial data–Participant follow-up Observed Summary Score from complete data)

In the following analyses, the Observed Summary Score estimate and the psychometric performance of the Observed Summary Score created from the complete simulated data were compared to the same parameters for each missingness scenario. This included assessments of estimate bias, convergent and known groups methods validity, test–retest reliability, and definition of meaningful change thresholds at the individual and group level. The description of these analyses is presented in supplementary material 3.

Results

Score estimate bias was found to increase with increasing number of days missing and depending on the missingness mechanism (Table 3). Here, the results are shown for 10% and 40% of the sample with missing data, with full results in the supplementary material 4. MNAR showed the most bias, followed by MAR then MCAR. In general, estimates with MCAR data did not show any meaningful bias, but showed increased variability (see Fig. 1). As such, future tables do not display MCAR but results are available in supplementary material 4. Estimates of MAR data had a bias of < 0.1 points in all but the most extreme cases of missingness (40% of

Table 3 Bias in mean summary score for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	2	3	4	5	6	7
MCAR	Low	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	0.000	0.000	− 0.001	− 0.001	− 0.001	0.001	0.003
	Medium	10	0.000	0.000	0.000	− 0.001	− 0.001	− 0.001	0.000
		40	0.000	0.000	− 0.001	− 0.001	− 0.001	− 0.003	− 0.001
	High	10	0.000	0.000	− 0.001	− 0.001	− 0.001	− 0.003	− 0.001
		40	0.000	0.000	0.000	0.000	0.001	0.001	0.003
MAR	Low	10	− 0.001	− 0.004	− 0.009	− 0.015	− 0.022	− 0.030	− 0.002
		40	− 0.006	− 0.016	− 0.035	− 0.060	− 0.090	− 0.121	− 0.016
	Medium	10	− 0.002	− 0.006	− 0.011	− 0.018	− 0.028	− 0.036	− 0.002
		40	− 0.007	− 0.022	− 0.044	− 0.074	− 0.112	− 0.145	− 0.015
	High	10	− 0.001	− 0.003	− 0.006	− 0.010	− 0.013	− 0.013	0.000
		40	− 0.003	− 0.011	− 0.020	− 0.034	− 0.049	− 0.049	− 0.007
MNAR	Low	10	− 0.005	− 0.010	− 0.017	− 0.030	− 0.053	− 0.090	− 0.066
		40	− 0.020	− 0.042	− 0.071	− 0.122	− 0.216	− 0.361	− 0.397
	Medium	10	− 0.007	− 0.015	− 0.026	− 0.043	− 0.071	− 0.117	− 0.052
		40	− 0.028	− 0.060	− 0.100	− 0.166	− 0.281	− 0.461	− 0.314
	High	10	− 0.010	− 0.020	− 0.034	− 0.057	− 0.096	− 0.158	− 0.030
		40	− 0.039	− 0.081	− 0.136	− 0.227	− 0.381	− 0.626	− 0.191

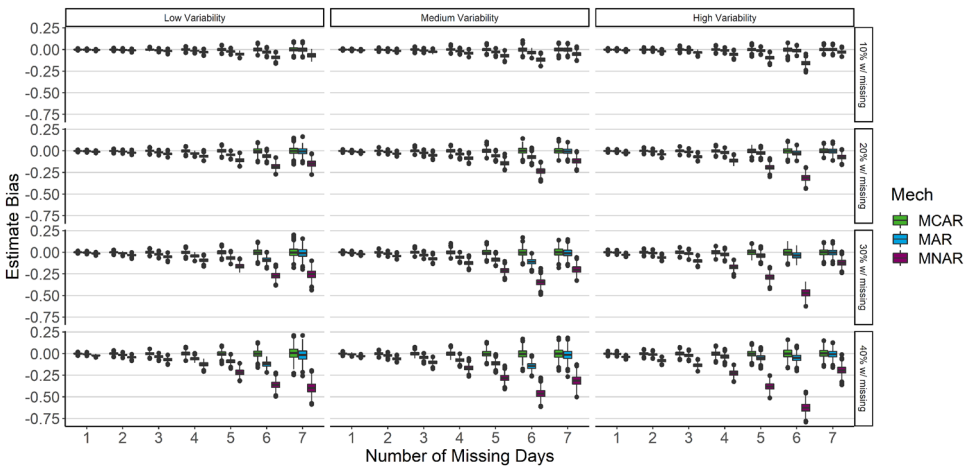


Fig. 1 Box plots showing the bias in the Observed Summary Score estimate compared to complete data. Boxes from left to right in each condition: MCAR, MAR, MNAR

patients missing 5 or 6 days of data). MNAR was severely impacted by missing data, with summary scores showing a bias of more than 0.5 points on the 0–10 scale in some scenarios. However, even MNAR data performed well when considering the typical 4/7 rule (maximum bias = 0.136 points). When compared to complete case analysis (i.e. all 7 days removed), scenarios with three days of missing data led to a similar level of bias as complete case analysis, but had the advantage of less variability (Fig. 1).

The psychometric properties showed similar results. For test–retest reliability, missing data led to less reliable results than complete case analysis, but these were mostly within 0.05 of the simulated ICC. The larger biases generally occurred in scenarios with 5 to 6 days of missingness and high variability. MNAR data were again a slight exception, with medium variability scenarios also showing ICC biases > 0.05 (Table 4). When patients had 4 days of missingness or fewer, the estimates had a negligible bias compared to the complete case analysis condition (7 day of missing data) and lower variability around these estimates. This led to a similar lower bound ICC estimate (Fig. 2).

In terms of validity, convergent validity results showed minimal bias in terms of correlation coefficient reduction (reduction in $r < 0.05$) in all but the most extreme scenarios (5 or 6 days of missingness for 40% of patients in high variability conditions; Table 5; Fig. 3). In general, this measurement property showed bias when missing data were present, but the magnitude of the bias was negligible in most cases. Known groups validity, however, was shown to underestimate the effect size of change across many different combinations of missing days, missingness mechanisms, and percent of sample with missing days. Table 6 and Fig. 4 display the percent bias of the known groups analysis effect size. This effect size is calculated between "moderate" and "severe" simulated patients and the bias is shown as the deviation from complete data. This table shows that there is up to 10% to 20% reduction in effect size of score difference between two known groups compared to complete data when using the typical 4/7-day rule (i.e. allowing 3 days of

Table 4 Bias in ICC for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	2	3	4	5	6	7
MAR	Low	10	0.000	− 0.001	− 0.002	− 0.003	− 0.006	− 0.010	0.000
		40	− 0.001	− 0.004	− 0.007	− 0.012	− 0.021	− 0.039	0.000
	Medium	10	− 0.001	− 0.002	− 0.003	− 0.006	− 0.010	− 0.018	− 0.001
		40	− 0.002	− 0.007	− 0.012	− 0.021	− 0.035	− 0.062	− 0.002
	High	10	− 0.001	− 0.003	− 0.006	− 0.011	− 0.018	− 0.035	0.000
		40	− 0.005	− 0.014	− 0.024	− 0.039	− 0.063	− 0.108	− 0.001
MNAR	Low	10	0.000	− 0.001	− 0.003	− 0.006	− 0.012	− 0.023	0.000
		40	− 0.001	− 0.005	− 0.012	− 0.023	− 0.041	− 0.075	− 0.011
	Medium	10	− 0.001	− 0.002	− 0.006	− 0.011	− 0.021	− 0.040	− 0.001
		40	− 0.002	− 0.009	− 0.021	− 0.040	− 0.069	− 0.118	− 0.010
	High	10	− 0.001	− 0.005	− 0.012	− 0.024	− 0.045	− 0.082	0.000
		40	− 0.005	− 0.019	− 0.042	− 0.075	− 0.125	− 0.195	− 0.006

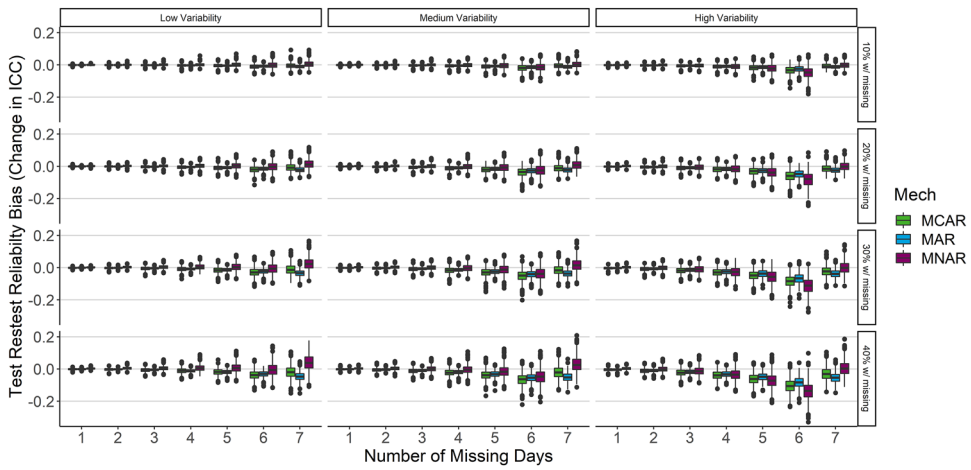


Fig. 2 Box plots showing the bias in the reliability of the Observed Summary Score estimate compared to complete data. This is shown in terms of ICC change. Boxes from left to right in each condition: MCAR, MAR, MNAR

Table 5 Bias in convergent validity correlation for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	**2**	**3**	**4**	**5**	**6**	**7**
MAR	Low	10	0.000	0.000	− 0.001	− 0.003	− 0.004	− 0.008	− 0.010
		40	− 0.001	− 0.002	− 0.005	− 0.011	− 0.019	− 0.032	− 0.047
	Medium	10	0.000	− 0.001	− 0.002	− 0.005	− 0.008	− 0.014	− 0.011
		40	− 0.001	− 0.004	− 0.009	− 0.019	− 0.032	− 0.055	− 0.050
	High	10	− 0.001	− 0.003	− 0.005	− 0.009	− 0.014	− 0.025	− 0.012
		40	− 0.004	− 0.009	− 0.018	− 0.031	− 0.049	− 0.083	− 0.053
MNAR	Low	10	0.001	0.001	0.001	0.001	0.001	− 0.001	0.006
		40	0.003	0.005	0.005	0.006	0.006	− 0.005	0.041
	Medium	10	0.001	0.001	0.000	− 0.001	− 0.004	− 0.014	0.003
		40	0.003	0.004	0.001	− 0.004	− 0.015	− 0.047	0.030
	High	10	0.001	0.000	− 0.004	− 0.010	− 0.021	− 0.048	− 0.001
		40	0.003	− 0.001	− 0.014	− 0.035	− 0.072	− 0.136	0.005

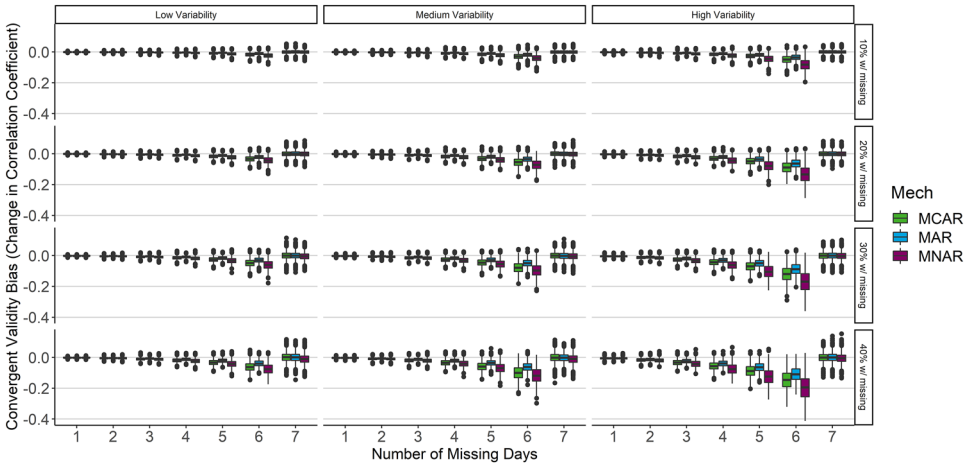


Fig. 3 Box plots showing the bias in the convergent validity correlations of the Observed Summary Score estimate compared to complete data. Boxes from left to right in each condition: MCAR, MAR, MNAR

Table 6 Percent bias in known groups effect size for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	2	3	4	5	6	7
MAR	Low	10	− 0.377	− 0.906	− 1.715	− 3.059	− 5.249	− 9.830	− 0.033
		40	− 1.371	− 3.434	− 6.385	− 10.560	− 16.712	− 27.260	− 0.152
	Medium	10	− 0.593	− 1.534	− 2.912	− 5.116	− 8.483	− 14.786	− 0.016
		40	− 2.323	− 5.745	− 10.053	− 16.062	− 24.469	− 36.450	0.272
	High	10	− 1.175	− 2.792	− 4.853	− 7.760	− 12.513	− 21.466	0.012
		40	− 4.236	− 9.586	− 15.770	− 23.226	− 32.912	− 46.331	0.018
MNAR	Low	10	− 0.455	− 1.620	− 3.587	− 6.816	− 12.049	− 20.852	0.539
		40	− 1.692	− 5.834	− 12.172	− 20.677	− 13.189	− 22.361	− 3.115
	Medium	10	− 0.649	− 2.505	− 5.684	− 10.497	− 17.790	− 28.776	0.545
		40	− 2.522	− 8.870	− 17.798	− 28.699	− 41.170	− 54.731	2.346
	High	10	− 1.150	− 4.186	− 8.983	− 16.194	− 26.503	− 40.099	0.189
		40	− 4.395	− 13.854	− 25.540	− 38.356	− 51.778	− 65.128	0.507

For low vs Moderate severity comparisons see supplementary material

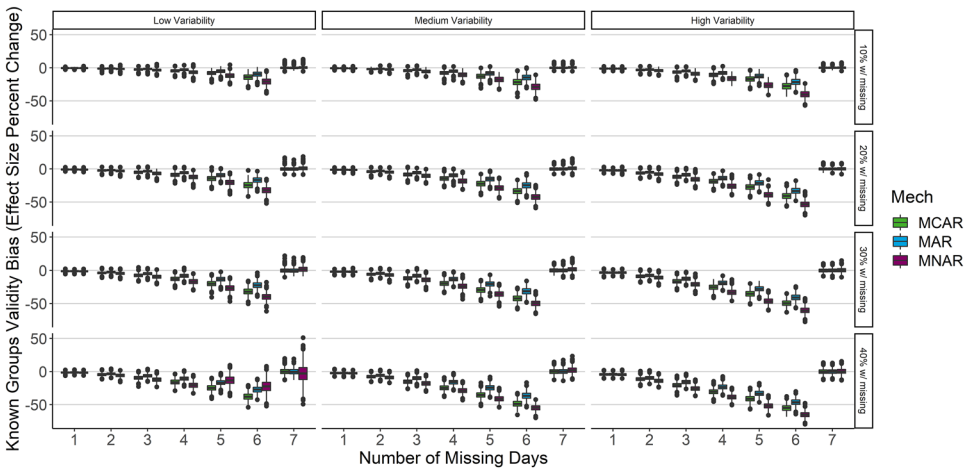


Fig. 4 Box plots showing the percent bias in the known groups validity of the Observed Summary Score estimate compared to complete data. This shows effect size differences between "moderate" and "severe" simulated groups. Boxes from left to right in each condition: MCAR, MAR, MNAR

data to be missing; Table 6, Fig. 4). This suggests that using known groups analysis as an assessment of validity when daily diary data are missing could severely underestimate the true difference between two known groups. Longitudinal validity (responsiveness to change) also showed bias across all mechanisms. The direction of the bias varied and was small in most cases (< 5%; Table 7). The MNAR mechanism, however, had larger biases, with the high variability condition showing > 10% bias in estimate in some cases. The variability around the estimate was large (Fig. 5).

With regard to interpretation thresholds, the minimal important difference was recovered well for all missingness mechanisms, with almost all scenarios showing < 5% bias in MID estimate compared to what would be seen if all data were available (Table 8). Interestingly, for MAR and MNAR scenarios, complete case analysis led to more bias than almost all scenarios which created a summary score from partial data. However, plots show that there is a lot of variability around MID estimates, meaning that any given scenario could potentially have a bias in the range of + 20% to − 20% percent (Fig. 6). Within-person responder definitions showed a complex pattern of results. Estimates from scenarios with the MCAR and MAR mechanism showed limited bias, whose direction was negative (i.e. underestimating the responder definition; Table 9; Fig. 7). However, MNAR mechanisms showed a negative bias with fewer missing days and a positive bias with more missing days. A complete case analysis (i.e. excluding patients with missing data, here the same as having all 7 days missing) was better able to return an unbiased responder definition.

Discussion

This study aimed to develop a resource for understanding the impact of missing daily diary data on score estimates, psychometric properties and interpretation thresholds. The tables and figures displayed in this work and the supplementary materials allow researchers to estimate the level of bias which could exist in their study either a priori (based

Table 7 Percent bias in ability to detect change effect size for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	**2**	**3**	**4**	**5**	**6**	7
MAR	Low	10	− 0.078	− 0.265	− 0.596	− 0.994	− 1.369	− 1.690	0.212
		40	− 0.351	− 1.125	− 2.426	− 4.030	− 5.792	− 7.320	1.488
	Medium	10	− 0.138	− 0.402	− 0.810	− 1.300	− 1.824	− 2.281	0.157
		40	− 0.561	− 1.556	− 3.264	− 5.300	− 7.343	− 8.972	1.314
	High	10	− 0.031	− 0.181	− 0.383	− 0.628	− 0.864	− 0.675	0.172
		40	− 0.183	− 0.756	− 1.580	− 2.545	− 3.249	− 3.199	0.800
MNAR	Low	10	− 0.237	− 0.499	− 0.839	− 1.400	− 2.387	− 3.909	− 0.287
		40	− 0.964	− 2.007	− 3.360	− 5.483	− 9.222	− 15.107	− 1.450
	Medium	10	− 0.363	− 0.765	− 1.278	− 2.047	− 3.290	− 5.166	− 0.234
		40	− 1.385	− 2.928	− 4.854	− 7.905	− 12.802	− 20.419	− 1.286
	High	10	− 0.522	− 1.097	− 1.858	− 3.033	− 4.869	− 7.626	− 0.230
		40	− 2.077	− 4.314	− 7.233	− 11.833	− 19.099	− 30.012	− 1.038

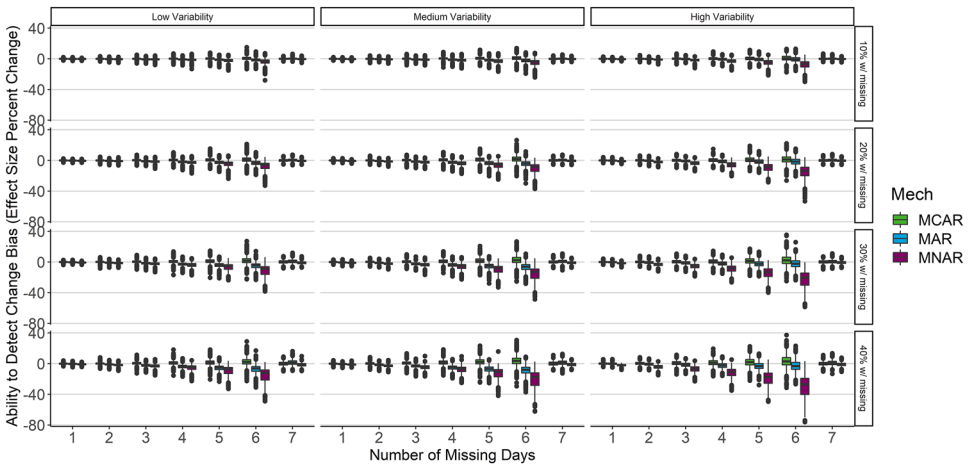


Fig. 5 Box plots showing the percent bias in the longitudinal validity (or ability to detect change) of the Observed Summary Score estimate compared to complete data. This shows the effect size of the change for participants who were simulated as changed on the PGIC between baseline and follow-up. Boxes from left to right in each condition: MCAR, MAR, MNAR

Table 8 Percent bias in minimal important difference threshold for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	2	3	4	5	6	7
MAR	Low	10	− 0.018	− 0.080	− 0.203	− 0.313	− 0.313	− 0.293	− 0.934
		40	− 0.086	− 0.365	− 0.862	− 1.321	− 1.615	− 1.816	− 4.924
	Medium	10	− 0.025	− 0.126	− 0.276	− 0.391	− 0.465	− 0.558	− 0.816
		40	− 0.186	− 0.491	− 1.178	− 1.742	− 1.918	− 1.987	− 4.374
	High	10	0.018	− 0.055	− 0.124	− 0.168	− 0.201	0.011	− 0.546
		40	− 0.033	− 0.265	− 0.619	− 0.860	− 0.766	− 0.736	− 3.450
MNAR	Low	10	0.007	0.006	0.017	0.103	0.281	0.599	1.031
		40	0.050	0.112	0.179	0.694	1.660	3.258	6.796
	Medium	10	0.011	0.022	0.037	0.110	0.403	0.850	0.877
		40	0.079	0.154	0.313	0.708	1.806	3.523	5.785
	High	10	− 0.002	− 0.022	− 0.067	− 0.018	0.201	0.610	0.567
		40	0.013	0.039	0.058	0.260	1.326	3.621	4.042

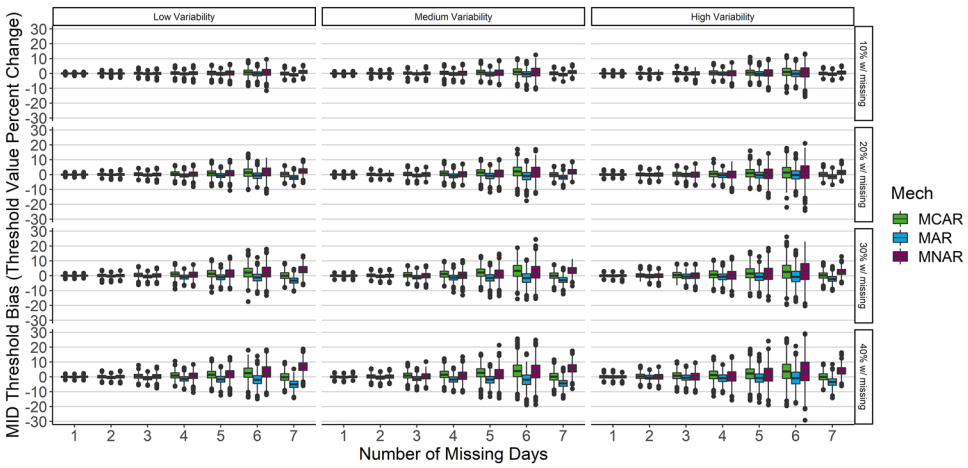


Fig. 6 Box plots showing the percent bias in the group-level minimal important difference estimate for the Observed Summary Score compared to complete data. Boxes from left to right in each condition: MCAR, MAR, MNAR

on previous work in the field) or after completing a data collection. This will allow them to assess the impact of bias when conducting their analyses either when using complete case analysis, using an existing missing data rule to create summary scores (e.g. patients must have at least 4 days out of 7) or when creating a new missing data rule. It is important to note that, while a missing data threshold may vary for different analysis, the purpose of assessing measurement

Table 9 Percent bias in responder definition for each missingness scenario

Mechanism	Variability	Percent	Number of missing days						
			1	2	3	4	5	6	7
MAR	Low	10	− 1.832	− 1.941	− 2.365	− 2.277	− 1.786	− 0.690	− 0.366
		40	− 5.937	− 7.705	− 7.681	− 7.820	− 8.653	− 4.272	− 0.860
	Medium	10	− 1.808	− 2.082	− 2.667	− 2.800	− 1.549	− 0.171	0.216
		40	− 6.344	− 8.290	− 8.416	− 8.571	− 7.504	− 2.135	− 1.304
	High	10	− 3.021	− 3.343	− 3.138	− 2.955	− 1.935	− 0.633	− 0.005
		40	− 8.992	− 10.937	− 11.690	− 11.350	− 9.841	− 3.660	− 1.345
MNAR	Low	10	− 0.710	− 1.526	− 1.424	− 0.925	− 0.254	1.144	0.668
		40	− 3.099	− 4.505	− 4.712	− 3.724	0.490	5.382	2.962
	Medium	10	− 1.107	− 1.752	− 1.704	− 0.962	0.163	1.254	0.304
		40	− 3.076	− 4.228	− 3.697	− 1.957	2.056	6.572	2.958
	High	10	− 1.464	− 1.424	− 0.848	− 0.471	0.655	1.029	0.202
		40	− 2.769	− 3.201	− 1.422	1.088	3.077	5.656	2.668

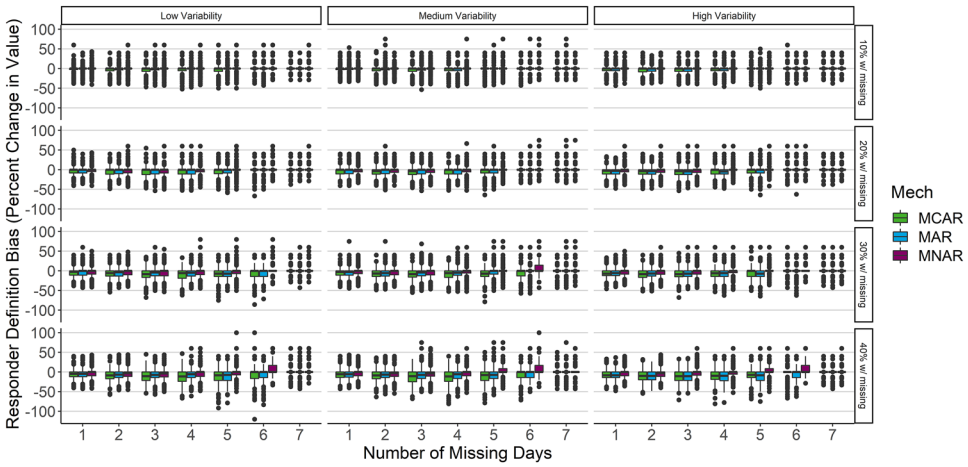


Fig. 7 Box plots showing the percent bias in the individual-level responder definition estimate for the Observed Summary Score compared to complete data. Boxes from left to right in each condition: MCAR, MAR, MNAR

properties is to have confidence that a score is valid and reliable. As such, the researcher should not alter the missing data rule in order to conduct the analyses, but instead understand the impact of their chosen missing data rule on their analyses.

This work has shown that the impact of missing data handling strategies varies depending on the analysis being conducted. For the majority of the measurement property assessments in the explored scenario, it was apparent that existing rules of thumb (creating a summary score for participants with 4 or more observed days of data) are comparable, if not preferable to removing these participants from the analysis (case wise deletion). Despite the bias in the psychometric parameters (especially when the data were explicitly MNAR), the precision of the estimates made them likely more reliable than when removing participants from the analysis. Therefore, this work supports a general rule of thumb that researchers can employ to derive summary scores for patients with at least 4 days of data in a 7-day period.

Some interesting results arose when assessing the meaningful change threshold. Similar to the psychometric results, the group-level minimal important difference thresholds (i.e. the point at which the difference between two treatment arms is meaningful) are likely more accurately defined by using all available data to create the summary score, at least when the data are MAR. However, when defining individual-level responder definitions, results show bias exists when making a summary score when data are missing. This bias (most often) leads to a negative skew in the responder definition which could lead to a smaller-than-appropriate responder definition. The consequence of this could be that participants are incorrectly classified as responders when they did not, in reality, respond. The impact of this is that when defining a meaningful change threshold, if this threshold is negatively biased, the results using this threshold will look *more* favourable than they should be. For a clinical trial, this could mean approving a treatment based on results showing that a treatment group met the meaningful change threshold, which if full data were available would have been higher.

Recommendation arising from this work varies depending on whether a MID or a responder definition is sought. For the MID, the scoring rule of at least 4 days out of 7 days can be used in the analysis, but the highest estimate (or rounding up the estimate) arising from the triangulation process should be used. For defining responder definitions using the ROC method, it is recommended that a sensitivity analysis is conducted, removing participants with missing data from the analysis prior to deriving a responder definition to understand how the definition is impacted. The analyses in this work showed that this approach led to minimal bias when using the ROC method to define responders.

Overall, the way the measurement properties of a summary score are assessed, interpreted and analysed should relate to the endpoint employed for the study. This does mean, however, that researchers should build their endpoint hierarchies with the knowledge that the chosen method to construct the summary score will impact their results, deviating in some way from the unobservable truth. Given the recent shift towards the estimand language [18], researchers should make sure that their objectives clearly indicate the construct they are estimating with their summary scores. The tables and figures in this work allow researchers to understand the effect that their chosen method of summary score creation will have on their results and ultimately consider robust derivation rules. As suggested above, researchers could also use alternative summary score derivation rules to be used as a sensitivity analysis in the case that they are deriving responder definitions, or in cases where the data may be MNAR (where the highest level of bias occurred). This will allow them to observe whether the psychometric properties hold under more realistic assumptions.

Researchers using this work should be aware of some of the limitations. For example, although the correlation matrices used simulated different levels of variability, these were somewhat arbitrary, and researchers should assess their own data structure when using the supplementary tables. Furthermore, the correlations between single days of data in the baseline and follow-up were mostly set to zero. This was a choice to simplify the analysis, but different results could be found if alternative relationships are specified.

To conclude, the current work suggests that the way a summary score is derived with regard to missing data has an impact on the score's measurement properties and interpretation. However, despite the clear message that missingness leads to bias, the impact is not uniform across all analyses explored here. Some analyses seem to tolerate missing data better than others. Furthermore, this work shows that the relationship between summary score missing data rules and analysis results is further moderated by the daily variability inherent in the condition, the number of participants with missing data, and the missingness mechanism underlying the incomplete data. The creation of summary scores is a form of data reduction. Specifically in the case of averaging over missing data, this is a form of imputation. As such, perhaps other methods of handling daily diary data are necessary, such as intensive longitudinal models where each daily diary observation can be, for example, nested within each week. Although this work allows researcher further insight into how to conduct research under the present paradigm of creating summary scores, it is perhaps instead time to start exploring methods which can model change across time using each individual datapoint available.

For researchers interested in exploring the scenarios presented in this work, an interactive tool has been created and is available in the supplementary material for download.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-022-03198-9.

Author Contributions PG conceived of the presented idea. PG and AW developed the theory and performed the computations. PG and AW verified the analytical methods. AW built the online tool. EB supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Funding This work was not funded, but was conducted while the authors were employees of Adelphi Values, UK.

Data availability Data underlying the simulation and the online tool are available from the corresponding author on request.

Code availability If code is required, requests can be made to the corresponding author and will be considered on a case for case basis.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Not required.

Consent to participate Not required.

Consent for publication Not required.

References

1. Okupa, A. Y., Sorkness, C. A., Mauger, D. T., Jackson, D. J., & Lemanske, R. F., Jr. (2013). Daily diaries vs retrospective questionnaires to assess asthma control and therapeutic responses in asthma clinical trials: Is participant burden worth the effort? *Chest, 143*(4), 993–999.
2. Lewandowski, A. S., Palermo, T. M., Kirchner, H. L., & Drotar, D. (2009). Comparing diary and retrospective reports of pain and activity restriction in children and adolescents with chronic pain conditions. *The Clinical journal of pain, 25*(4), 299.
3. Bennett, A. V., Amtmann, D., Diehr, P., & Patrick, D. L. (2012). Comparison of 7-day recall and daily diary reports of COPD symptoms and impacts. *Value in Health, 15*(3), 466–474.

4. Coons, S. J., Eremenco, S., Lundy, J. J., O'Donohoe, P., O'Gorman, H., & Malizia, W. (2015). Capturing patient-reported outcome (PRO) data electronically: The past, present, and promise of ePRO measurement in clinical trials. *The Patient-Patient-Centered Outcomes Research, 8*(4), 301–309.
5. Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient non-compliance with paper diaries. *BMJ, 324*(7347), 1193–1194.
6. Tiplady, B. (2010). Diary Design Considerations. EPro: Electronic Solutions for Patient-reported Data, 167.
7. Bingham, C. O., Gaich, C. L., DeLozier, A. M., Engstrom, K. D., Naegeli, A. N., de Bono, S., Banerjee, P., & Taylor, P. C. (2019). Use of daily electronic patient-reported outcome (PRO) diaries in randomized controlled trials for rheumatoid arthritis: Rationale and implementation. *Trials, 20*(1), 1–8.
8. Strunk, R. C., Bender, B., Young, D. A., Sagel, S., Glynn, E., Caesar, M., & Lawhon, C. (2002). Predictors of protocol adherence in a pediatric asthma clinical trial. *Journal of allergy and clinical immunology, 110*(4), 596–602.
9. Holzbaur, E., & Ross, J. (2014). Risks, impacts, and mitigation of missing epro data on clinical trials. *Value in Health, 17*(3), A204.
10. Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials*. CRC Press.
11. Coens, C., Pe, M., Dueck, A. C., Sloan, J., Basch, E., Calvert, M., Campbell, A., Cleeland, C., Cocks, K., Collette, L., Devlin, N., Dorme, L., Flechtner, H. H., Gotay, C., Griebsch, I., Groenvold, M., King, M., Kluetz, P. G., Koller, M., … Bottomley, A. (2020). International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: Recommendations of the SISAQOL Consortium. *The Lancet Oncology, 21*(2), e83–e96.
12. Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical methods in medical research, 23*(5), 440–459.
13. Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual review of psychology, 54*(1), 579–616.
14. Griffiths, P., Floden, L. L., & Hudgens, S. (2017, October). Scoring and interpretation of daily diary data in the presence of non-ignorable missing data. In QUALITY OF LIFE RESEARCH (Vol. 26, No. 1, pp. 70–70). VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS: SPRINGER.
15. Griffiths, P., Floden, L., Doll, H., Morris, M., & Hudgens, S. (2018, October). Psychometric properties in the face of missing data-a simulation study assessing the effect of missing data on test-retest reliability in diary studies. In QUALITY OF LIFE RESEARCH (Vol. 27, pp. S55-S55). VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDS: SPRINGER.
16. Fayers, P. M., & Machin, D. (2013). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes*. John Wiley & Sons.
17. Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis, 22*(3), 329–343.
18. Lawrance, R., Degtyarev, E., Griffiths, P., Trask, P., Lau, H., D'Alessio, D., Griebsch, I., Wallenstein, G., Cocks, K., & Rufibach, K. (2020). What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *Journal of Patient-Reported Outcomes, 4*(1), 1–8.