Check for updates

Received: 20 April 2023 | Revised: 21 October 2023 | Accepted: 26 October 2023

DOI: 10.1111/cea.14441

ORIGINAL ARTICLE

WILEY

Evaluation of measurement errors in the Patient-Oriented Eczema Measure (POEM) outcome

Ariane Duverdier[1,2,3] | Guillem Hurault[1] | Kim S. Thomas[4] | Adnan Custovic[5] | Reiko J. Tanaka[1]

[1]Department of Bioengineering, Imperial College London, London, UK

[2]UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK

[3]Department of Computing, Imperial College London, London, UK

[4]Centre for Evidence Based Dermatology, School of Medicine, University of Nottingham, Nottingham, UK

[5]National Heart and Lung Institute, Imperial College London, London, UK

Correspondence
Reiko J. Tanaka, Department of Bioengineering, Imperial College London, London, UK.
Email: r.tanaka@imperial.ac.uk

Funding information
The UKRI CDT in AI for Healthcare http://ai4health.io (EP/S023283/1) and the British Skin Foundation (005/R/18).

Abstract

Background: The Patient-Oriented Eczema Measure (POEM) is the recommended core outcome instrument for atopic dermatitis (AD) symptoms. POEM is reported by recalling the presence/absence of seven symptoms in the last 7 days.

Objective: To evaluate measurement errors in POEM recordings due to imperfect recall.

Methods: Using data from a clinical trial of 247 AD patients aged 12–65 years, we analysed the reported POEM score (r-POEM) and the POEM derived from the corresponding daily scores for the same seven symptoms without weekly recall (d-POEM). We quantified recall error by comparing the r-POEM and d-POEM for 777 patient-weeks collected from 207 patients, and estimated two components of recall error: (1) recall bias due to systematic errors in measurements and (2) recall noise due to random errors in measurements, using a bespoke statistical model.

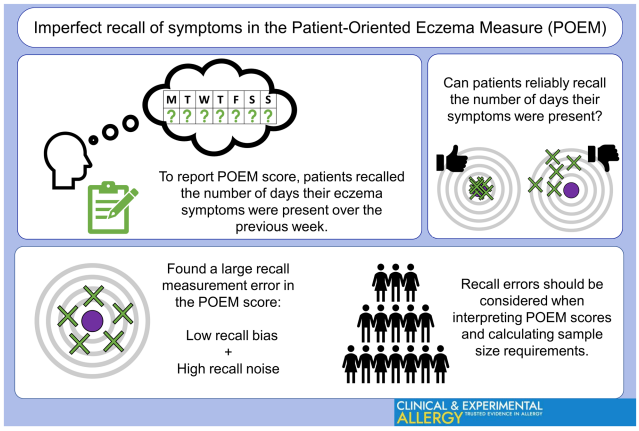
Results: POEM scores have a relatively low recall bias, but a high recall noise. Recall bias was estimated at 1.2 points lower for the r-POEM on average than the d-POEM, with a recall noise of 5.7 points. For example, a patient with a recall-free POEM of 11 (moderate) could report their POEM score anywhere from 5 to 14 (with 95% probability) because of recall error. Model estimates suggested that patients tend to recall itch and dryness more often than experienced (positive bias of less than 1 day), but less often for the other symptoms (bleeding, cracking, flaking, oozing/weeping and sleep disturbance; negative bias ranging 1–4 days).

Conclusions: In this clinical trial data set, we found that patients tended to slightly underestimate their symptoms when reporting POEM, with significant variation in how well they were able to recall the frequency of their symptoms every time they reported POEM. A large recall noise should be taken into consideration when interpreting POEM scores.

KEYWORDS
atopic dermatitis, measurement error, POEM, recall bias, recall noise

Ariane Duverdier and Guillem Hurault contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Authors. *Clinical & Experimental Allergy* published by John Wiley & Sons Ltd.



GRAPHICAL ABSTRACT

The Patient-Oriented Eczema Measure (POEM) is a recommended score to assess eczema symptoms from a patient's perspective with a weekly recall. Patients tended to slightly underestimate their symptoms when reporting POEM with significant recall noise. Recall errors should be considered when interpreting POEM scores. More research is needed to evaluate methods to reduce poor recall (such as aide-memoirs).

1 | INTRODUCTION

Atopic dermatitis (AD) is the most common chronic skin condition, with a considerable adverse impact on individual patients and healthcare systems.[1–4] AD is characterised by multiple objective physical signs (such as cracking and erythema) and subjective symptoms (such as itching and sleep disturbance). Several scores are used to measure the severity of AD and monitor changes in patients' disease state. The Harmonising Outcome Measures for Eczema (HOME) core outcome set for clinical trials of AD initiative[5] recommends using the Patient-Oriented Eczema Measure (POEM) and peak Numerical Rating Scale 11 for itch intensity over 24 h (NRS itch) to measure patient symptoms, the Eczema Area and Severity Index (EASI) to measure clinical signs, the Dermatology Life Quality Index questionnaires for adults, children, and infants (DLQI, CDLQI and IDQoL respectively) to measure quality of life and Recap of atopic eczema (RECAP) or AD Control Tool (ADCT) to assess long-term control. These scores help clinical decision-making, the evaluation of the effects of interventions in clinical trials,[6,7] and can be useful for predicting the evolution of AD severity.[8] It is therefore important to assess the properties of the scores, including their validity (whether the score measures what it is supposed to measure), reliability (whether the score is free from measurement errors), interpretability (whether the score has a qualitative meaning) and responsiveness to change (whether the score can detect change in severity over time).[9,10] These properties can be related to each other. For example, a high level of measurement errors could impact the validity and interpretability of a score. In this context, measurement error refers to the difference between a measured quantity and its 'true' value and can be divided into two components: bias due to systematic errors in measurements and noise due to random errors (Figure 1).[11] Bias refers to the average error between the measurement and its true value and noise refers to the variability/dispersion of errors, as opposed to the variance around the bias estimate.[11]

Key messages

- POEM can be subject to large measurement errors, due to the imperfect recall of symptoms.
- Trial participants tended to slightly underestimate their symptoms when reporting POEM with significant recall noise.
- Recall errors should be considered when interpreting POEM scores and calculating sample size requirements.

Six of the eight instruments (POEM, DLQI, CDLQI, IDQoL, RECAP and ADCT) listed above rely on weekly recall. When selecting the recall period of a patient-reported outcome (PRO) measure, it is important to ensure that patients/carers are able to recall symptoms easily and accurately.[12] The Food and Drug Administration (FDA) reported that 'PRO instruments that call for patients to rely on memory, especially if they must […] average their response over a period of time, are likely to undermine content validity'.[13] Moreover, errors in the recall of symptoms may not be easily detectable if we do not have access to more frequently measured symptom data, such as those derived from daily diaries. Memory biases (cognitive heuristics), such as the peak-end rule (peak and recency effects on memory),[14] likely play a role in the recall error. The length of recall periods may influence how patients interpret questions and select relevant information for responses.[15,16] For instance, when reporting the frequency of anger symptoms, patients reported less severe and more frequent episodes when asked to recall symptoms over 1 week compared to 1 year.[17] Studies on weekly recall found that patients underestimate the frequency[18,19] and overestimate the intensity of their symptoms[15,18]; on average, patients had a harder time recalling the frequency of their symptoms than their intensity.[18]

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

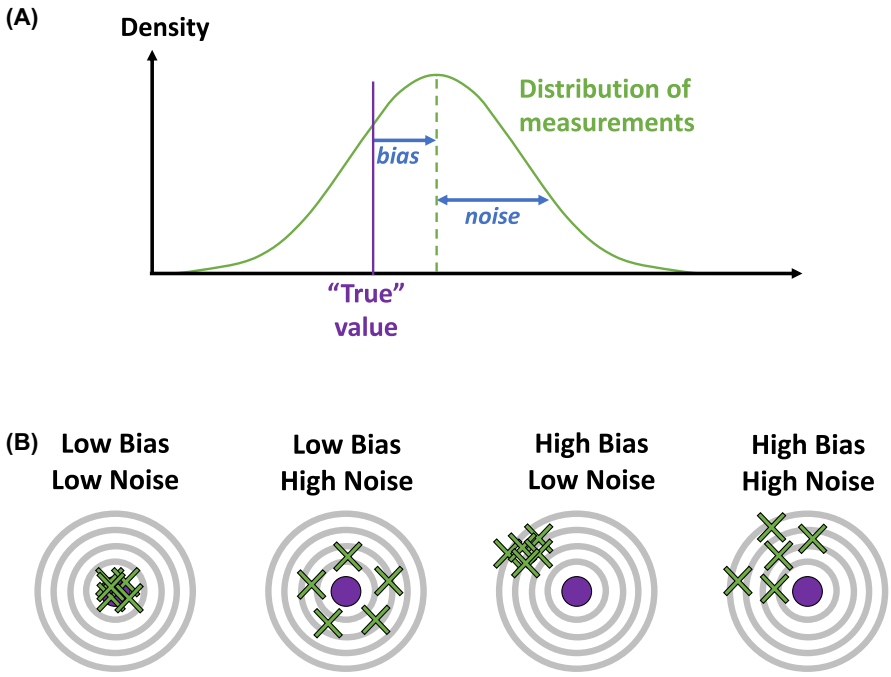


FIGURE 1 Illustration of measurement error. (A) Definition of bias and noise. (B) Schematic for high/low bias and noise. Bias refers to the average error between the measurement and its true value, and noise refers to the variability/dispersion of errors. Noise is a measure of the variability of recall errors, not the variance around the bias estimate.

In this study, we focus on POEM. The POEM score summarises the presence of seven symptoms over the past week.[20] Assessment requires patients (or caregivers) to recall the absence/presence of symptoms over the past 7 days. While POEM is widely used as a recommended score in practice, a systematic review identified that research on measurement errors in POEM has been inconclusive[21]; the two studies the review identified were of insufficient quality to interpret evidence of measurement error. More recently, a study aiming to predict AD severity found that POEM was subject to more measurement error than EASI and the objective component of the SCORing AD index (oSCORAD),[22] which are based on physical examination at a single time point. To address this research gap in the literature, we carried out a series of analyses to quantify the recall error (measurement error due to imperfect recall of symptoms) in the reporting of the POEM score. We analysed data collected previously in a clinical trial of topical AD treatment which included weekly POEM scores and the corresponding daily scores for the seven symptoms of POEM, thereby providing a unique opportunity to compare the reported POEM susceptible to recall error with the POEM obtained from daily symptoms, without weekly recall.

2 | METHODS

2.1 | Data sources

In this study, we used data collected in a randomised trial sponsored by GlaxoSmithKline under Parexel project #225866 for dose-finding for a topical AD treatment. The trial recruited 247 patients aged 12–65 years with mild to severe AD in Japan, USA and Canada.

The data we used included up to six measurements of POEM per patient and symptom scores, based on the content of POEM, recorded in daily electronic diaries over 16 weeks. During the trial, patients were asked to indicate the severity of the six AD symptoms (itchy skin, bleeding skin, oozing skin, cracked skin, flaky skin and dry or rough skin) in the past 24 h on a discrete scale ranging from 0 ('Absent') to 10 ('Worst imaginable'). Additionally, patients were asked to measure 'sleep impact' daily, on a scale of 0–10. The data was pre-processed according to the steps outlined in Supplementary Methods S1 in Data S1, Figures S1 and S2.

For the study we report in this article, the data were grouped into patient-weeks. Each patient-week record contains a POEM score and the matching daily diary entries from the same week for a specific patient and a specific week.

2.2 | Definition of r-POEM and d-POEM: deriving POEM score without weekly recall from daily symptom scores

Here, we use the term 'POEM' to refer to the instrument itself (the set of questions patients are asked to complete to measure the POEM score) rather than its measured values. The POEM score is obtained from the self-assessment of seven symptoms (bleeding, cracking, dryness, flaking, itching, oozing/weeping and sleep disturbance). Patients/carers are asked how many days each symptom occurred in the past week. Their answers are graded on a discrete scale from 0 to 4 (0 = 'no days', 1='1 or 2 days', 2='3 or 4 days', 3='5 or 6 days' and 4 = '7 days'). The graded answers for the seven symptoms are finally summed to produce a score taking discrete values between 0 and 28.

We use the term r-POEM to refer to the POEM score reported (recalled) by the patients in the study. r-POEM scores were measured by asking patients to recall their symptoms over the past

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

week. To quantify recall error in the POEM score, we introduce d-POEM, which refers to the POEM score derived from the daily symptom scores' diaries without weekly recall. d-POEM scores were calculated from the daily symptom scores, using four steps outlined by the POEM definition (Figure 2). Firstly, the daily symptom scores (0–10) were dichotomised into absence or presence (0 or 1) by a threshold. The daily symptom scores represent the intensity of the symptom, while POEM records the frequency of symptoms; we used a threshold to convert the daily intensity scores to daily absence or presence of symptoms. If the daily symptom score was greater than or equal to the threshold, the symptom was considered present, otherwise absent (Figure 2A). We set the threshold to be 2 to minimise the sum of squared difference between r-POEM and d-POEM scores (Figure S3A). Secondly, the number of days when each symptom was present was counted over a 7-day window (Figure 2B). Thirdly, the number of present days (0–7) for each symptom was graded (0–4) according to the POEM definition (Figure 2C). Finally, d-POEM score was obtained as the sum of the scores for each symptom (Figure 2D).

2.3 | Analysis

2.3.1 | Imputation of missing absence/presence of daily symptoms

Calculation of d-POEM scores requires daily symptom recordings for the 7 previous days. Our data allowed us to calculate d-POEM scores only for 202 (out of 845) patient-weeks because 30% of daily symptoms recordings were missing (12,271 out of 41,405) (Figure S2). To avoid discarding 76% (643 out of 845) of r-POEM scores that did not have the corresponding d-POEM score, we imputed missing absence/presence of daily symptoms using Markov chain models (MCM, Supplementary Methods S2 in Data S1), with a prediction horizon of ≤7 days. The models were defined in the EczemaPred R package[23] for each symptom and fit on the full data set of daily symptom scores (17,398 daily scores for each symptom from 247 patients). A weakly informative prior was set on the transition matrix of the MCM, such that a patient's severity is more likely to transition to adjacent severity scores (e.g. 4–5) than to non-adjacent scores

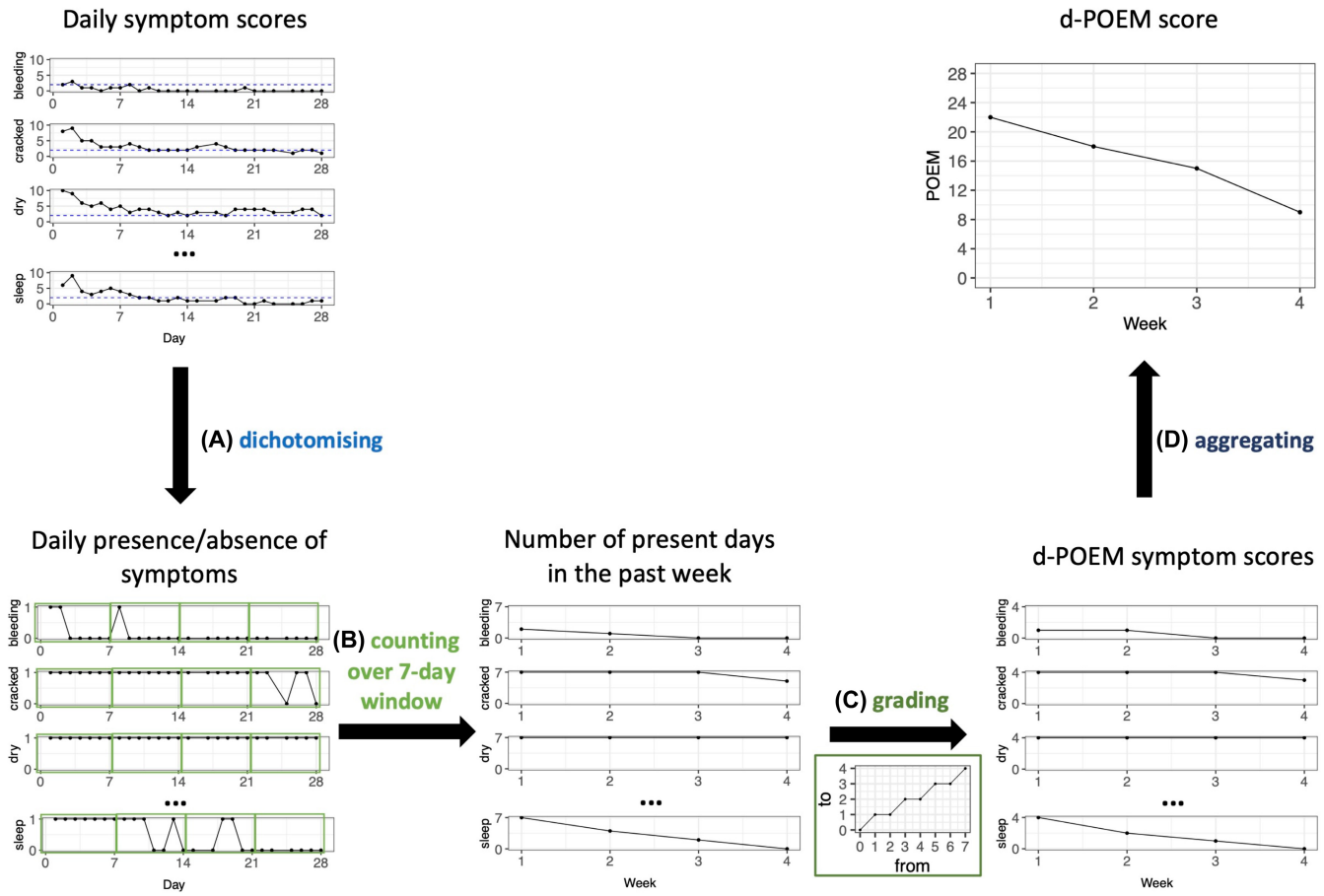


FIGURE 2 Derivation of d-POEM scores (Patient-Oriented Eczema Measure scores without weekly recall) from the daily symptoms scores. The data are shown for a representative patient. (A) The daily symptom scores are dichotomised into daily presence (1 when daily score ≥ threshold) or absence (0 otherwise). (B) The number of days each symptom is present is counted over a 7 days window. (C) The number of present days for each symptom is graded according to the POEM definition (0 = 'no days', 1 = '1 or 2 days', 2 = '3 or 4 days', 3 = '5 or 6 days' and 4 = '7 days') to obtain d-POEM symptom scores. D) The d-POEM symptom scores are aggregated to obtain the d-POEM score.

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(e.g. 1–10) (Figure S4). The performance of the imputation models was evaluated by 10-fold cross-validation stratified by patients on the full data set. We evaluated the ability of the models to predict absence/presence (the 0/1 classification accuracy) and to discriminate between the two classes (the area under the receiving operating characteristic curve, AUROC).

2.3.2 | Quantification of recall error

We quantified recall error (measurement error due to imperfect recall of symptoms) as the difference between r-POEM (recalled POEM) and d-POEM (POEM without weekly recall) scores. We estimated recall bias and recall noise by calculating the average and standard deviation, respectively, of the difference between r-POEM and d-POEM scores (r-POEM − d-POEM). We also computed recall bias and recall noise for each symptom to understand how recall errors can affect each symptom, because estimates for the aggregate POEM could be confounded by the prevalence of symptoms (e.g. low error if a symptom is less prevalent).

To complement estimates of recall error by the method above, we developed a bespoke statistical machine learning model to estimate recall bias and recall noise in the number of days each symptom was present in the week (Figure 3). We refer to this model as the recalled days model. The added benefit of using the recalled days model is that it provides estimates of recall error on a more interpretable scale (days) and is less affected by grading. The grading, that is, grouping of days in the POEM symptom scores (1='1 or 2 days', 2='3 or 4 days' etc.), may account for and buffer some of the recall error thereby improving the reliability of the POEM score. For example, even if a patient unreliably recalls 1 day with dry skin symptoms instead of 2, the 1 day difference is masked by the grading. To fully understand the ability of patients to recall symptoms, without the masking effects by grading, we estimated recall error in the number of days recalled. The recalled days model was used to infer recall bias and recall noise estimates for each symptom by comparing the number of days with symptoms present according to daily diaries and the recalled number of days inferred from the r-POEM symptom score. For the n-th reported POEM symptom score, the model assumes that the number of days ($\hat{d}_n$) a patient recalled experiencing the symptom (inferred from the r-POEM symptom score) is distributed around the number of days (d_n) the patient experienced the symptom (according to daily diaries) with a recall bias and noise (Figure 3, Supplementary Methods S3 in Data S1).

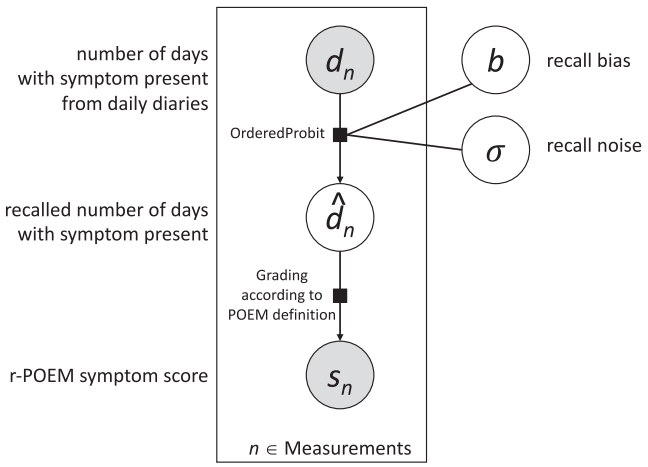


FIGURE 3 Diagram of the recalled days model illustrated as a factor graph, where grey and white circles represent measured variables and latent (unmeasured) variables estimated by the model respectively. The model was fit to each of the seven POEM symptoms separately. The recalled number of days with symptom present is graded according to the POEM definition (0 = 'no days', 1 = '1 or 2 days', 2 = '3 or 4 days', 3 = '5 or 6 days' and 4 = '7 days') to obtain the r-POEM (recalled Patient-Oriented Eczema Measure) symptom scores.

2.3.3 | Fitting of statistical machine learning models

We fitted the MCM for imputation of missing absence/presence of daily symptoms and the statistical machine learning model (recalled days model) for estimation of recall error using Bayesian inference in the probabilistic programming language Stan.[24] We used the Hamiltonian Monte Carlo algorithm with four chains of 4000 iterations each, including a 50% warm-up. We found no evidence of an absence of convergence by monitoring trace plots and R-hat statistics.

3 | RESULTS

3.1 | Imputation of missing absence/presence of daily symptoms

The absence/presence of daily symptoms was imputed for 9842 missing values (26% of daily symptoms used in the analysis), enabling us to use data from 777 patient-weeks (collected from 207 patients) (Figure S2). The average classification accuracy across cross-validation folds for 1-day-ahead predictions (most frequent prediction horizon with 14,088/17,136 test observations) ranged from 90.8% to 94.8% and the AUROC from 0.94 to 0.96 for the seven symptoms (Figure S5). The average accuracy decreased with increasing prediction horizon but remained greater than 81.4% for all symptoms, even for 7-day-ahead predictions (29 test observations).

3.2 | Quantification of recall error

The r-POEM score was 1.2 points lower than d-POEM score on average (Figure 4A). Negative recall biases (patients reported lower r-POEM than d-POEM) were observed in most of the symptoms; recall biases remained fairly low, ranging from −0.28 to 0.05 across symptoms (Figure 4B). Recall noise was 5.7 for r-POEM scores (Figure 4C).

To provide more interpretable estimates of recall error, we estimated the errors in terms of the number of days each symptom was present using the recalled days model. The estimated recall biases

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

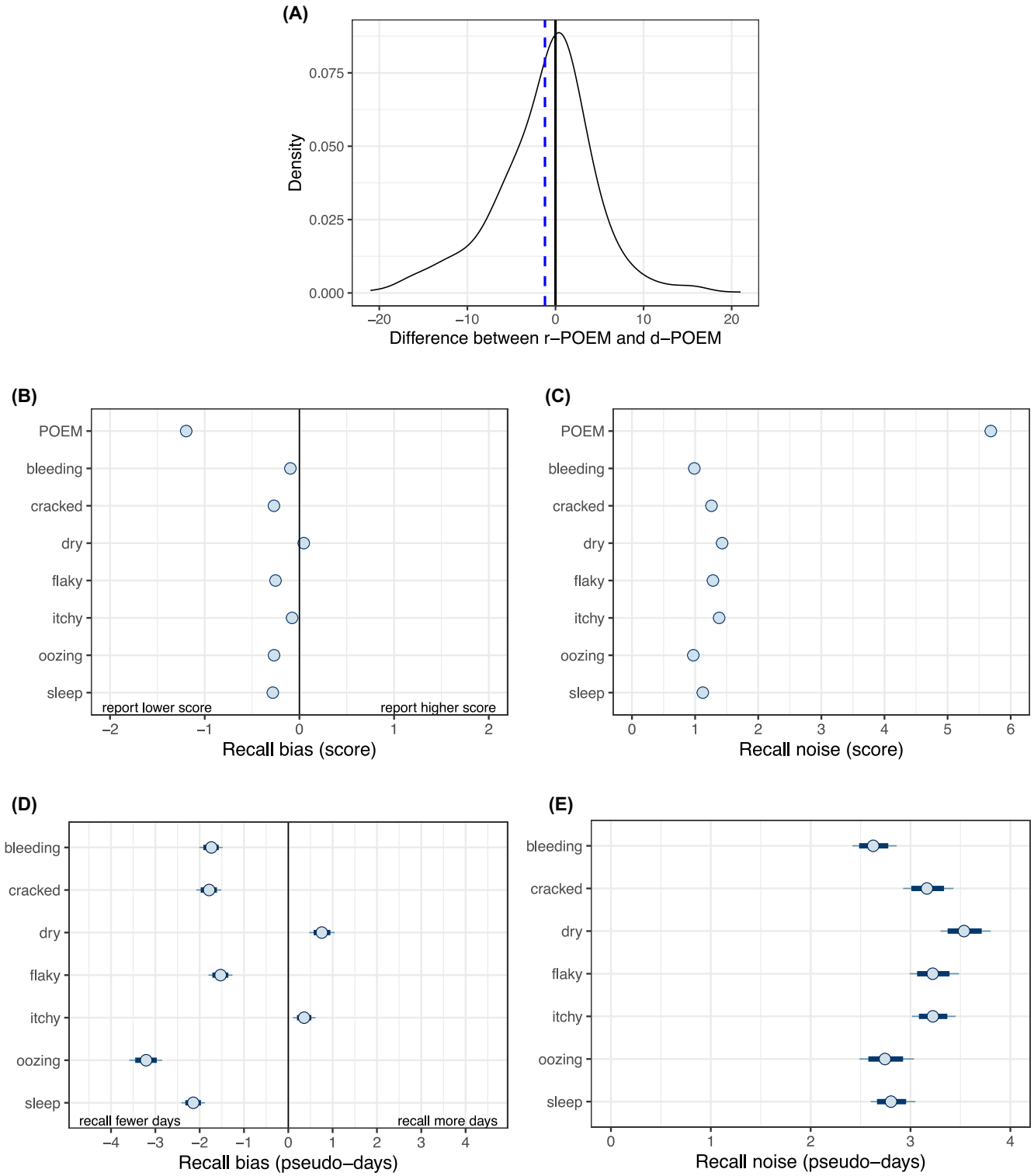


FIGURE 4 Recall error in POEM and its seven symptoms. (A) Distribution of the pairwise difference between r-POEM and d-POEM (r-POEM – d-POEM); the dotted line indicates the average difference. (B–E) Recall bias and recall noise for the POEM score and its seven symptom scores (B,C) and those in the number of days each of the symptoms was experienced (D,E). r-POEM is the recalled POEM score, and d-POEM is the POEM score derived from the daily diaries without weekly recall. Average (B) and standard deviation (C) of the pairwise difference between r-POEM and d-POEM scores. Estimates (D,E) were obtained from the recalled days model (Supplementary Methods S3 in Data S1); the inner and outer intervals are 95% and 80% credible intervals. A negative recall bias indicates patients recalled fewer days with symptoms than those according to their daily symptom scores. There is no recall error estimate for the POEM score (D,E) because the recalled number of days is defined only at the symptom level as answered in the POEM questionnaire.

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

varied between symptoms and were positive or negative (Figure 4D). For example, patients tend to recall the presence of itch and dryness more often than they experienced (positive bias of less than 1 day), but recall fewer days for other five symptoms (negative bias ranging between 1 and 4 days across the symptoms). Recall noise was estimated to be around 3 days for all seven symptoms (Figure 4E).

We conducted a sensitivity analysis on the threshold values used to calculate d-POEM scores (Figures S3 and S6, Supplementary Results S4 in Data S1). The threshold (=2) used in this study results in the smallest recall error in the POEM score (Figure S3A) and the recall bias being closest to 0 (Figure S6A). The recall noise varied from 4.5 to 6.2, across the threshold values (Figure S6C).

To illustrate the effect of recall bias and recall noise on the measurement of r-POEM, we estimated the distribution of r-POEM score of three representative patients (Table S1) with mild (d-POEM score = 5), moderate (d-POEM = 11) and severe (d-POEM = 21) AD, using the recalled days model (Figure 5). A patient with mild d-POEM could report an r-POEM score anywhere between 2 and 11 (with 95% probability), with an average r-POEM score of 6.4 (Figure 5A). A patient with moderate d-POEM could report an r-POEM between 5 and 14 (with 95% probability), with an average r-POEM score of 10.0 (Figure 5B). A patient with severe d-POEM could report an r-POEM between 9 and 20 (with 95% probability), with an average of 14.9 (Figure 5C).

We also conducted a complete case analysis without imputation and confirmed that our results were not sensitive to the imputation of missing absence/presence of daily symptoms (Supplementary Results S5 in Data S1, Figures S7 and S8).

4 | DISCUSSION

Herein, we reported the first detailed investigation of recall error in the measurement of POEM. We analysed data collected from 247 patients with mild to severe AD and demonstrated evidence of recall error in POEM; the patients tended to underestimate their symptoms when reporting POEM, recalling fewer days with symptoms than they experienced, with significant recall noise.

The extent of recall error is likely to vary for different cohorts of patients; we found recall error to vary between symptoms and symptoms' severity and patients with different patterns of symptoms may exhibit different recall biases. Symptoms that are more bothersome may be more accurately recalled or even slightly overestimated. For instance, itching and dryness have the smallest recall biases, both of which are slightly positive. Less prevalent symptoms, like oozing, have larger recall biases and are generally underestimated.

4.1 | Implications for clinical trials

PROs (collected daily or weekly) allow collection of data between clinic visits and can be a useful way to capture data at multiple timepoints with minimal burden in chronic-relapsing conditions such as AD. Recall bias may have little impact on the interpretation of data collected with weekly recall in randomised controlled trials, since well-matched groups would be equally affected. However, a large recall noise may impact the interpretation of results. A larger measurement noise means the need for a larger sample size to detect a given effect size.[25–27] Future trials should consider recall noise when performing power analysis and calculation of sample size requirements because measurement noise reduces the power of a clinical trial.[25,27] Low power reduces the chance of detecting a true effect and the chance that a statistically significant result reflects a true effect.[28] This means that low-powered studies may falsely claim no treatment effect (false negative), especially if looking for small effects,[25] or report a statistically significant result that does not represent a true treatment effect (false positive).[28] It is important to consider a large recall noise in POEM when interpreting results from previous clinical trials using POEM.

4.2 | Strengths and limitations

We demonstrated that missing value imputation can be easily integrated in an analytical workflow using simple yet accurate models, while missing values may be seen as an obstacle to working with

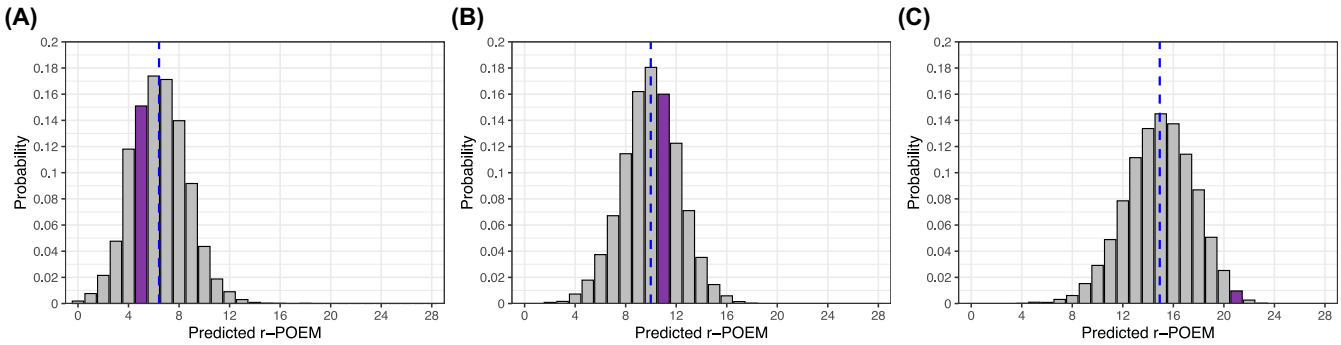


FIGURE 5 Distribution of r-POEM predicted by the recalled days model for three representative patients with (A) mild (d-POEM of 5), (B) moderate (d-POEM of 11) and (C) severe atopic dermatitis (d-POEM of 21). The purple bar and dotted blue line indicate the d-POEM score and the average predicted r-POEM respectively. A patient was randomly selected from each of mild, moderate and severe AD patients in the data set. r-POEM is the recalled POEM score, and d-POEM is the POEM score derived from the daily diaries without weekly recall.

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

daily recordings. We implemented statistical models (Markov chain models) to impute missing absence/presence of daily symptoms with high accuracy. It allowed us to triple the number of patient-weeks used in the analysis. To ensure the high accuracy, we imputed only missing daily symptoms with a prediction horizon of 7 days or less. We confirmed that our results on recall error were not sensitive to the missing data imputation.

We also designed a bespoke statistical machine learning model to quantify recall error in the number of days patients recalled the presence of each symptom. The model provides estimates that are less affected by the grading, allowing for a better understanding of the recall ability of patients, and on a more interpretable scale (days). For example, the conventional approach (calculating the recall error by r-POEM − d-POEM) indicated that sleep had a recall bias of approximately −0.3. Using the recalled days model, we found that sleep had a recall bias of approximately −2 days, meaning that patients tend to recall symptoms 2 days less than they experienced according to their daily diaries.

We used a unique data set with weekly- and daily-reported AD symptoms collected prospectively. However, the two outcomes were not collected in the same way. POEM records the frequency of symptoms, while the daily data were based on the intensity of symptoms. To calculate the recall-free frequency of symptoms, we dichotomised the intensity of daily symptoms into a binary absence/presence variable; this dichotomisation could have contributed to measurement error. Dichotomising intensity is inherent to the POEM score: When patients report their scores, they decide what they consider as absence/presence of symptoms. In this study, we assumed a single constant threshold for absence/presence of symptoms for all symptoms and patients; this is a pragmatic assumption as any patient- or time-specific scoring would make any quantitative analysis difficult. The sensitivity analysis showed that using different thresholds would not affect our conclusion that POEM can be subject to large recall errors. Dichotomising intensity to absence/presence in previous studies[18,29] also often used the same threshold across symptoms.

The recall error estimated in this study is a combination of the actual recall error and any potential error from the dichotomisation process. We chose a threshold that minimises the difference between r-POEM and d-POEM scores. Future research could investigate how patients determine absence/presence of symptoms to understand how the POEM should be interpreted. For example, a study on how patients interpret frequency questions has demonstrated that short recall periods encourage the reporting of more minor events than longer periods.[17] In our data set, recall error was smallest with a threshold of 2, rather than a threshold of 1; a higher threshold may compensate for the reporting of more minor events in the 24 h recall period of the daily diaries than the week of POEM.

While we demonstrated that patient-reported POEM has a large noise due to imperfect recall, it is unclear how patients actually complete the POEM questionnaire. The measurement process may be more complex than simply recalling the frequency of a symptom around the 'true' frequency, shifted by a constant bias. For example, responses could be disproportionally influenced by recent symptoms. Unfortunately, we could not find conclusive evidence of recency bias due to multicollinearity in the daily symptom recordings (scores are correlated with previous scores) and a low signal-to-noise ratio in the data. More research is needed to evaluate the recency bias in POEM. We did not estimate patient-specific recall bias and noise in a multilevel model due to the insufficient number of observations per patient (up to four), although recall error could also be patient-dependent. Other types of measurement errors beyond the recall error quantified in this study could also be investigated.

5 | CONCLUSION

We have shown that POEM can be subject to a large recall noise, but a fairly low recall bias. More research is required to determine the impact of measurement errors on other weekly reported outcome instruments and to evaluate interventions (such as the use of aide-memoires) aimed at reducing recall error.

AUTHOR CONTRIBUTIONS

AD: Conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualisation and writing—original draft. GH: Conceptualisation, formal analysis, investigation, methodology, software, validation, visualisation and writing—original draft. KST and AC: Validation and writing—review and editing. RJT: Conceptualisation, funding acquisition, project administration, resources, supervision, validation, writing—original draft, and writing—review and editing.

ACKNOWLEDGEMENTS

This work was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (EP/S023283/1) and the British Skin Foundation (005/R/18). The data analysed in this study were obtained from a clinical trial sponsored by GlaxoSmithKline.

FUNDING INFORMATION

This work was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (EP/S023283/1) and the British Skin Foundation (005/R/18).

CONFLICT OF INTEREST STATEMENT

KST works at the Centre of Evidence Based Dermatology, where the POEM outcome instrument was developed and is a member of the Executive Group for the HOME core outcome set initiative.

DATA AVAILABILITY STATEMENT

All code for the analysis is available at https://github.com/arianeduverdier/POEM-recall-error

ORCID

Ariane Duverdier https://orcid.org/0000-0003-0132-3525
Guillem Hurault https://orcid.org/0000-0002-1052-3564

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kim S. Thomas https://orcid.org/0000-0001-7785-7465
Adnan Custovic https://orcid.org/0000-0001-5218-7071
Reiko J. Tanaka https://orcid.org/0000-0002-0769-9382

REFERENCES

1. Lusignan S, Alexander H, Broderick C, et al. Patterns and trends in eczema management in UK primary care (2009–2018): a population-based cohort study. *Clin Exp Allergy*. 2021;51:483-494.
2. Langan SM, Mulick AR, Rutter CE, et al. Trends in eczema prevalence in children and adolescents: A global asthma network phase I study. *Clin Exp Allergy*. 2023;53:337-352.
3. Lusignan S, Alexander H, Broderick C, et al. The epidemiology of eczema in children and adults in England: a population-based study using primary care data. *Clin Exp Allergy*. 2021;51:471-482.
4. Langan SM, Irvine AD, Weidinger S. Atopic dermatitis. *Lancet*. 2020;396:345-360.
5. Williams HC, Schmitt J, Thomas KS, et al. The HOME Core outcome set for clinical trials of atopic dermatitis. *J Allergy Clin Immunol*. 2022;149:1899-1911.
6. Chopra R, Silverberg JI. Assessing the severity of atopic dermatitis in clinical trials and practice. *Clin Dermatol*. 2018;36:606-615.
7. Gooderham MJ, Bissonnette R, Grewal P, Lansang P, Papp KA, Hong CH. Approach to the assessment and management of adult patients with atopic dermatitis: a consensus document. Section II: tools for assessing the severity of atopic dermatitis. *J Cutan Med Surg*. 2018;22:10S-16S.
8. Hurault G, Dominguez-Huttinger E, Langan SM, Williams HC, Tanaka RJ. Personalized prediction of daily eczema severity scores using a mechanistic machine learning model. *Clin Exp Allergy*. 2020;50:1258-1266.
9. Mokkink LB, Prinsen CAC, Bouter LM, Vet HCW, Terwee CB. The consensus-based standards for the selection of health measurement iNstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther*. 2016;20:105-113.
10. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. 2018;27:1159-1170.
11. Hibbert DB. Systematic errors in analytical measurement results. *J Chromatogr A*. 2007;1158:25-32.
12. Norquist JM, Girman C, Fehnel S, DeMuro-Mercon C, Santanello N. Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration. *Qual Life Res*. 2012;21:1013-1020.
13. U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for industry patient-reported outcome measures use in medical product development to support labeling claims. 2009. Agency/Docket Number: Docket No. FDA-2006-D-0362; Document Number:E9-29273.(Federal Register Volume 74, Number 235, Pages 65132-65133). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims
14. Fredrickson BL, Kahneman D. Duration neglect in retrospective evaluations of affective episodes. *J Pers Soc Psychol*. 1993;65:45-55.
15. Peasgood T, Caruana JM, Mukuria C. Systematic review of the effect of a one-day versus seven-day recall duration on patient reported outcome measures (PROMs). *Patient*. 2023;16:201-221.
16. Schwarz N. Self-reports: how the questions shape the answers. *Am Psychol*. 1999;54:93-105.
17. Winkielman P, Knäuper B, Schwarz N. Looking back at anger: reference periods change the interpretation of emotion frequency questions. *J Pers Soc Psychol*. 1998;75:719-728.
18. Rydén A, Leavy OC, Halling K, Stone AA. Comparison of daily versus weekly recording of gastroesophageal reflux disease symptoms in patients with a partial response to proton pump inhibitor therapy. *Value Health*. 2016;19:829-833.
19. Beutler S, Daniels J, Laddis A. Do patients underestimate their symptoms in hindsight? An ambulatory assessment study on the frequency of dissociation in posttraumatic stress disorder. *Front Psychother Trauma Dissociat*. 2020;4:105-120.
20. Charman CR, Venn AJ, Williams HC. The patient-oriented eczema measure. *Arch Dermatol*. 2004;140:1513-1519.
21. Gerbens LAA, Prinsen CAC, Chalmers JR, et al. Evaluation of the measurement properties of symptom measurement instruments for atopic eczema: a systematic review. *Allergy*. 2017;72:146-163.
22. Hurault G, Roekevisch E, Schram ME, et al. Can serum biomarkers predict the outcome of systemic immunosuppressive therapy in adult atopic dermatitis patients? *Skin Health Dis*. 2022;2(1):e77.
23. Hurault G, Stalder JF, Mery S, et al. EczemaPred: a computational framework for personalised prediction of eczema severity dynamics. *Clin Transl Allergy*. 2022;12:e12140.
24. Carpenter B, Gelman A, Hoffman MD, et al. Stan: a probabilistic programming language. *J Stat Softw*. 2017;76:1-32.
25. Philips GW, Jiang T. Measurement error and equating error in power analysis. *Pract Assess Res Eval*. 2016;21(9):1-12.
26. Kobak KA, Kane JM, Thase ME, Nierenberg AA. Why do clinical trials fail? *J Clin Psychopharmacol*. 2007;27:1-5.
27. Sjoding MW, Cooke CR, Iwashyna TJ, Hofer TP. Acute respiratory distress syndrome measurement error. potential effect on clinical study results. *Ann Am Thorac Soc*. 2016;13:1123-1128.
28. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365-376.
29. Schneider S, Stone AA. Distinguishing between frequency and intensity of health-related symptoms from diary assessments. *J Psychosom Res*. 2014;77:205-212.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Duverdier A, Hurault G, Thomas KS, Custovic A, Tanaka RJ. Evaluation of measurement errors in the Patient-Oriented Eczema Measure (POEM) outcome. *Clin Exp Allergy*. 2024;54:207-215. doi:10.1111/cea.14441

13652222, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cea.14441 by Imperial College London, Wiley Online Library on [12/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License