

Supplementary File 1. Missing Data Imputations.

Phendo is an observational research app and participants do not receive prompts from the research team to track any given item at a certain time. They are free to track (or not track) any given item as they wish. Consequently, missingness in the data occurs due to a variety of possible reasons that are not always known or easy to distinguish. For example, a period not tracked for a day could mean that the participant did not have a period, or they chose not to track, or did not use the app at all that day. To circumvent this issue, we took several measures. First, we limited data to days for which the participant tracked their pain, exercise and menstrual status at least once, as a proxy for app use. Next, we assigned a score of zero for pain on days where the participant had tracked exercise or menstrual status but not pain. This approach is motivated by 2 reasons. First, the nature of the pain question in Phendo (i.e., “Where is the pain?”, “How severe is the pain?”) assumes the participants to track when they feel pain and therefore a “No Pain” response is neither available in the app nor would make sense. Second, multiple imputation methods impute such that the resulting imputations are limited to the observed values and distributions. Thus by default it would omit the possibility of a zero in the resultant pain score distribution, which increases risk of overestimation of the scores in the sample.

BMI (calculated from participant reported height and weight) and education level were missing for 22% and 19% of the participants, respectively, and menstrual status was missing (i.e., not tracked) 22% of the time in the dataset. We imputed these 3 variables using multivariate imputations by chained equations [1] according to the heteroscedastic linear two-level structure of the data (i.e., hierarchical where, participant is the clustering variable) following standard multilevel multiple imputation methods. [1-4] We used two-level predictive mean matching for BMI and education level, which is a semi-parametric imputation method that limits imputations

to the observed values and can preserve non-linear relations in the observed data, therefore the imputations do not deviate from the observed distribution[5] and two-level logistic regression for imputing menstrual status, using the rest of the dataset as the predictors. As per published recommendations,[1, 2] we also included the raw pain variable (i.e., with the missing values) as a predictor, to account for the possibility of an association between the missingness pattern of pain to these imputed variables. To assess the plausibility of the imputations and any significant deviance from the structure of the raw, non-imputed data, we inspected the imputation convergence plots, distributions of the imputed variables which are provided in Supplementary Figures 3 and 4.

Model specification. We used a zero-inflated negative binomial (ZINB) distribution when modeling the total pain outcome, as it has been demonstrated to provide the best fit for outcomes with over-dispersion and zero-inflation.[6-8] ZINB models consider two sources of zero observations: “sampling zeros” that are part of the underlying sampling distribution (i.e., negative binomial) and “structural zeros” that cannot score anything other than zero (i.e., participant did not track).[6] This virtue of the ZINB models allows for specification of the imputed zeros and prevents the risk of over-estimating effects and generates more conservative estimates for predictors of interest by estimating a separate zero-inflation term, as well as conditional model.[6] We specified the zero-inflation term such that it was dependent on the exercise variable for the day, in addition to specifying an overall general zero-inflation structure in the outcome through inclusion of an intercept, based on recommendations. [8] Menstrual status was not a significant predictor of zero-inflation and therefore removed from the zero-inflation term during the modeling process. We included participants who had at least 11 pairs of consecutive days of data in the final analytic sample as this provided sufficient amount of data to

1) ensure model convergence and improve reliability and accuracy of the estimates, particularly the random effects and their variances[9-12], and 2) adequately infer participants' habitual weekly exercise frequency by considering at least three weeks' worth of tracking to compute the weekly exercise frequency.

Supplementary File 2. Imputation Model diagnostics.

Appropriateness and plausibility of the estimates from imputed models were inspected following published guidelines. First, we used measures of missing data information to assess pooled estimate variances. The fraction of missing information (λ) is interpreted as the proportion of variation in the parameter of interest due to the missing data. The relative increase in variance due to nonresponse (r) is interpreted as the proportional increase in the sampling variance of the parameter of interest that is due to the missing data. Values of λ over 0.5 indicate that the influence of the imputation model on the results is larger than that of the complete-data model, suggesting potential problems in the imputations. Supplementary Table 1 provides results of these variance estimates, indicating satisfactory imputation and model fit.

Supplementary Table 1. Measures of Missing data information

Conditional Fixed Effects	Total Pain Score		Difference in Pain	
	λ	r	λ	r
Intercept	0.21	0.27	0.23	0.31
Menstrual Status	0.13	0.15	0.19	0.23
Previous Day Pain	0.01	0.01	0.00	0.00
Body Mass Index	0.13	0.15	0.23	0.31
Mean Weekly Exercise Frequency	0.00	0.00	0.01	0.01
Previous Day exercise	0.01	0.01	0.00	0.00
Some College Education Level	0.26	0.36	0.35	0.55
College or Higher Education Level	0.23	0.31	0.21	0.28
Mean Weekly Exercise Frequency * Previous Day exercise	0.00	0.00	0.00	0.00
Zero Inflation Terms				
Intercept	0.00	0.00	0.00	0.00
Same Day Exercise	0.00	0.00	0.00	0.00

Next, we inspected propensity scores, which is a more recent and increasingly accepted method for inspecting the suitability of data imputation.[2, 13, 14] The goal is to compare the distributions of observed and imputed data conditional on the missingness probability. Under the missing at random (MAR) assumption, the conditional distributions of the observed and missing data should be similar if the assumed model for creating multiple imputations has a good fit. To do this, we first estimate the probability of each record being incomplete (i.e., “response propensity”) in the presence of missing data by conditioning on the response indicators as well as the observed covariates. The probabilities are then averaged over the imputed datasets to obtain stability. Supplementary Figure 3 plots BMI, education category

and menstrual status against the propensity score in each dataset. The distributions of the blue and red points are match up well without significant discrepancies (e.g., mismatch in patterns, imputed data systematically shifted toward one side of the axis).

Supplementary Table 2. Post-hoc analyses with endometriosis diagnosis included as a covariate. Conditional model results of the negative binomial model estimation of day-level total pain score (N=608).

Random Effects	Variance (95% CI)	
Participant (Intercept)	1.10 (0.99, 1.22)	

Fixed Effects	Log Odds (SE)	z-score
Intercept	1.37*** (0.12)	10.97
Menstrual Status	0.25*** (0.01)	21.40
Previous day Pain	0.02*** (0.01)	21.40
Body Mass Index	0.01* (0.004)	2.81
Mean weekly Exercise Frequency	-0.06** (0.02)	-3.01
Previous day exercise	0.09** (0.02)	3.85
Clinician diagnosis of endometriosis	-0.07 (0.10)	0.01
Self-diagnosis of endometriosis	-0.11 (0.11)	-1.01
Some college education level	0.22 (0.13)	-1.63
College or higher education level	-0.01 (0.12)	-0.12
Mean weekly Exercise Frequency*Previous day exercise	-0.03*** (0.01)	-3.42

SE=Standard Error. *p=0.001, ** p <0.001, ***p<0.0001. B coefficients are rate ratios. BMI =Body Mass Index. BMI and previous day pain were group mean centered.

Supplementary Table 3. Post-hoc analyses with endometriosis diagnosis included as a covariate. Conditional model results of the regression model estimation of pain score difference (N=1009).

Conditional Random Effects	Variance (95% CI)	
Participant (Intercept)	13.34 (12.09, 14.93)	

Fixed Effects	B coefficient (SE)	z-score
Intercept	2.45*** (0.46)	5.22
Menstrual status	1.46*** (0.08)	16.98
Previous day pain	-0.86*** (0.01)	-144.11
Body mass index	0.07* (0.01)	4.47
Mean weekly exercise frequency	-0.27** (0.09)	-3.03
Previous day exercise	0.92*** (0.18)	5.13
Clinician diagnosis of endometriosis	-0.05 (0.32)	-0.16
Self-diagnosis of endometriosis	-0.45 (0.43)	-1.29
Some college education level	-0.30 (0.51)	-0.58
College or higher education level	-1.72** (0.47)	-3.67
Mean weekly exercise frequency*Previous day exercise	-0.14* (0.06)	-2.31

SE=Standard Error. *p<0.05, ** p <0.01, ***p<0.0001. Body Mass Index and previous day pain were group mean centered.

References

1. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*; Vol 1, Issue 3 (2011). 2011.
2. Van Buuren S. *Flexible imputation of missing data*: CRC press; 2018.
3. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*. 2018;21(1):111-49.
4. Barnard J, Rubin DB. Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*. 1999;86(4):948-55.
5. Little RJ. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*. 1988;6(3):287-96.

6. Hu M-C, Pavlicova M, Nunes EV. Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *Am J Drug Alcohol Abuse*. 2011;37(5):367-75. doi: 10.3109/00952990.2011.597280. PubMed PMID: 21854279.
7. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*. 2017;9(2):378-400.
8. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. Modeling zero-inflated count data with glmmTMB. *bioRxiv*. 2017:132753. doi: 10.1101/132753.
9. Schunck R. Cluster Size and Aggregated Level 2 Variables in Multilevel Models. A Cautionary Note. 2016. 2016;10(1). Epub 2016-07-20. doi: 10.12758/mda.2016.005.
10. Bell B, Ferron J, Kromrey J, editors. Cluster Size in Multilevel Models: The Impact of Sparse Data Structures on Point and Interval Estimates in Two-Level Models 2008.
11. Austin PC, Leckie G. The effect of number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation*. 2018;88(16):3151-63. doi: 10.1080/00949655.2018.1504945.
12. Snijders TAB. Power and sample size in multilevel modeling. In: Everitt B, Howell D, editors. *Encyclopedia of Statistics in Behavioral Science*. 3: Wiley; 2006. p. 1570–3.
13. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. *Emerging Themes in Epidemiology*. 2017;14(1):8. doi: 10.1186/s12982-017-0062-6.
14. Bondarenko I, Raghunathan T. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*. 2016;35(17):3007-20. doi: <https://doi.org/10.1002/sim.6926>.