

# skipTrack: An R package for Identifying Skips in Self-Tracked Mobile Menstrual Cycle Data

Luke Duttweiler <sup>1</sup>, Shruthi Mahalingaiah <sup>2</sup>, and Brent Coull <sup>1</sup>




<sup>1</sup> Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States

<sup>2</sup> Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States

✉ Corresponding author

DOI: [10.21105/joss.06928](https://doi.org/10.21105/joss.06928)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Sehrish Kanwal](#)  

## Reviewers:

- [@ankurankan](#)
- [@nilabhrardas](#)

Submitted: 03 June 2024

Published: 16 September 2024

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

Mobile apps that allow users to self-track menstrual cycle lengths and symptoms are now widely available and frequently used (Fox & Duggan, 2010). Multiple studies (consider (Bull et al., 2019; Li et al., 2020; Mahalingaiah et al., 2022)) have taken advantage of these uniquely large data sets to gain insight on characteristics of the menstrual cycle, which is an important vital sign (Diaz et al., 2006). Due to the self-reported nature of the gathered data, recorded cycle lengths may be inflated if users skip tracking any cycle related bleeding events in the app. A non-trivial number of incorrectly inflated cycle lengths in a data set will be damaging to the reliability and reproducibility of analysis results.

Current solutions to this problem of non-adherence (skipped tracking) in cycle length reporting include removing implausibly long cycles that exhibit no user-app interaction (Li et al., 2020), identifying possibly inaccurate cycles based on user-specific average cycle lengths (Li et al., 2022), or *ad hoc* removal of cycles based on well-established menstrual cycle characteristics such as average cycle length or cycle length difference. The skipTrack package implements a Bayesian hierarchical model that is the first to explicitly use information on both an individual's cycle length **and** regularity to identify errors in recorded cycle lengths that arise from user non-adherence in logging one or more bleeding events.

## Statement of Need

Analyses involving large amount of user-tracked menstrual cycle data sets are becoming more prevalent. Identifying recorded cycle lengths that result from skips in tracking one or more period bleeding events (hereafter referred to as 'skipped cycles') is crucially important for maintaining the validity of these studies. The skipTrack package provides easy to use software in R that can identify skipped cycles in menstrual cycle data based on a pre-specified Bayesian hierarchical model. The resulting inference on possible skipped cycles may then be included by a researcher *a priori* in an analysis, or may be used to develop a multiple-imputation scheme.

Additionally, while based on the Bayesian hierarchical model from (Li et al., 2022), the model used by skipTrack includes components for both cycle length mean and regularity. This allows the model to correctly adjust for individuals with irregular cycles who are often excluded from menstrual cycle analyses, despite the important information their data contains. Finally, the skipTrack model and software lead to many possible useful extensions including the addition of regression models for both cycle length mean and regularity, an auto-regressive modeling structure for sequential cycle lengths from the same individual, and a method for the inclusion of user-app interaction or other external data to help with skip identification. These updates, along with open availability and ease-of-use, will provide researchers easy access to high level modeling techniques for mobile menstrual cycle data.

## The SkipTrack Model

We present a short overview of the SkipTrack model and notation here.

Let  $y_{ij}$  be the  $j$ th recorded cycle length provided by participant  $i$ . We assume that

$$y_{ij} \sim \text{LogNormal}(\mu_i + \log(c_{ij}), \tau_i)$$

where  $\mu_i$  is the natural log of individual  $i$ 's cycle length median,  $\tau_i$  is the precision of the distribution (providing a measure of regularity), and  $c_{ij}$  is an integer-valued parameter that represents the number of **true** cycles occurring in recorded cycle  $y_{ij}$ . For example, if  $c_{ij} = 1$ , then  $y_{ij}$  is a true cycle length, if  $c_{ij} = 2$  then  $y_{ij}$  is the length of two true cycles added together, and so on.

Then we assume,

$$\mu_i \sim \text{Normal}(\mu, \rho) \qquad \tau_i \sim \text{Gamma}(\theta, \phi)$$

where the natural log of  $\mu$  gives the overall population cycle length median,  $\rho$  is a precision parameter,  $\theta$  is the mean of the Gamma distribution and  $\phi$  is the rate.

Finally,

$$c_{ij} \sim \text{Categorical}(\pi_1, \pi_2, \dots, \pi_K)$$

where  $\pi_k = \Pr(c_{ij} = k)$  and  $K$  is the maximum number of skips allowed in the model.

## Package Description

The skipTrack package contains tools for fitting the SkipTrack model, visualizing model results, diagnosing model convergence, and simulating example data.

- **Model Fitting:** In order to fit the SkipTrack model, the code employs a Markov Chain Monte Carlo (MCMC) algorithm composed of Gibbs sampling steps. Model fitting may be accessed through the `skipTrack.fit()` function, and is accomplished through this easy-to-use interface that allows users to select the number of MCMC chains to run, the number of iterations to run per chain, and the parameters used to initialize each chain.
- **Visualizing Results:** Model results may be visualized or retrieved through standard reporting and visualization functions (`summary()`, `plot()`, etc.).
- **Diagnosing Convergence:** MCMC convergence diagnostics (traceplots, effective sample size, and the Gelman-Rubin potential scale reduction factor) are multivariate and multi-chain and are provided using the R package `genMCMCDiag` (Duttweiler, 2024), accessible through `skipTrack.diagnostics()`.
- **Simulating Data:** Data simulation options using `skipTrack.simulate()` are included which allow a user to simulate example data from the SkipTrack model, the generative model provided in Li et al. (2022), or a provided mixture model.

## Availability

A stable version of skipTrack is available on CRAN, and a development version is publicly available on GitHub (<https://github.com/LukeDuttweiler/skipTrack>).

## Acknowledgements

Research reported in this publication was supported by the National Institute of Environmental Health Sciences (NIEHS) Grants T32ES007142, P30 ES000002, and R01 ES035106. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## References

- Bull, J. R., Rowland, S. P., Scherwitzl, E. B., Scherwitzl, R., Danielsson, K. G., & Harper, J. (2019). Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *NPJ Digital Medicine*, 2(1), 83. <https://doi.org/10.1038/s41746-019-0152-7>
- Diaz, A., Laufer, M. R., Breech, L. L., & others. (2006). Menstruation in girls and adolescents: Using the menstrual cycle as a vital sign. *Pediatrics*, 118(5), 2245–2250. <https://doi.org/10.1542/peds.2015-4480>
- Duttweiler, L. (2024). *genMCMCDiag: Generalized convergence diagnostics for difficult MCMC algorithms*. <https://doi.org/10.32614/cran.package.genmcmcdiag>
- Fox, S., & Duggan, M. (2010). *Mobile health 2010*. Pew Internet & American Life Project Washington, DC.
- Li, K., Urteaga, I., Shea, A., Vitzthum, V. J., Wiggins, C. H., & Elhadad, N. (2022). A predictive model for next cycle start date that accounts for adherence in menstrual self-tracking. *Journal of the American Medical Informatics Association*, 29(1), 3–11. <https://doi.org/10.1093/jamia/ocab182>
- Li, K., Urteaga, I., Wiggins, C. H., Druet, A., Shea, A., Vitzthum, V. J., & Elhadad, N. (2020). Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile-health data. *NPJ Digital Medicine*, 3(1), 79. <https://doi.org/10.1038/s41746-020-0269-8>
- Mahalingaiah, S., Fruh, V., Rodriguez, E., Konanki, S. C., Onnela, J.-P., Figueiredo Veiga, A. de, Lyons, G., Ahmed, R., Li, H., Gallagher, N., & others. (2022). Design and methods of the apple women's health study: A digital longitudinal cohort study. *American Journal of Obstetrics and Gynecology*, 226(4), 545–e1. <https://doi.org/10.1016/j.ajog.2021.09.041>