

---

# NOF1-BCD - a framework for causal estimates with N-of-1 Bayesian Digital Twins

---

Anonymous Authors<sup>1</sup>

## Abstract

Single-person self-experimental data, so-called N-of-1 trials, are the gold standard for inference on individual treatment effects; however, they often face data scarcity, limiting inference on effect estimates. Also, they are not always feasible, and often, only observational data is available. Here, we propose NOF1-BCD, a framework for individual-level causal inference for treatment effects that provides regularization through priors in a Bayesian Digital Twin. First, we leverage large language models to efficiently synthesize external data sources—such as population-level studies and domain knowledge—into informative priors. Second, these priors are used in a Bayesian Digital Twin applied to individual-level data. Third, the Digital Twin serves as a prediction model to estimate the Average Period Treatment Effect, a counterfactual contrast for each individual. We characterize the framework and demonstrate it using smartwatch-collected physical activity and sleep recovery data, assessing the potential of LLM-elicited priors to improve estimation efficiency without compromising causal inference. Applications of the framework can extend beyond healthcare.

## 1. Introduction

Estimating how a treatment affects a specific individual—rather than an average patient—remains a core challenge in personalized medicine. Digital Twins (DTs) offer a promising approach: computational models of individual patients that continuously integrate incoming data and can simulate counterfactual outcomes (Emmert-Streib et al., 2025; Vallée, 2023). Recent work has begun combining DTs with Large Language Models (LLMs), leveraging LLMs’ capacity to

---

<sup>1</sup>Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

synthesize vast amounts of medical knowledge (Sun et al., 2024; Yang et al., 2025; Makarov et al., 2025).

For truly personalized causal inference, the study design of N-of-1 trials remains the gold standard (Nikles & Mitchell, 2015; Piccininni et al., 2024). A single individual alternates between treatment and control periods, generating time-series data that supports causal analysis at the individual level (Daza, 2018; Konigorski et al., 2025). Here, periods denote time segments in which the treatment assignment is held constant. Recently, Daza et al. (2025) introduced Model-Twin Randomization (MoTR), a framework for constructing causal Digital Twins from observational N-of-1 wearable data to estimate individual treatment effects. However, N-of-1 studies face a fundamental limitation: data scarcity. External knowledge— from prior studies, clinical guidelines, or domain expertise— could help inform and regularize estimates, but manually synthesizing this information is time-consuming and requires specialized expertise.

We propose NOF1-BCD (N-of-1 Bayesian Causal Digital Twin), a framework that addresses this gap by using LLMs to efficiently elicit informative priors from external sources, which then regularize a Bayesian Digital Twin for individual-level causal inference. Our approach has three components:

1. **LLM-based elicitation of priors from published literature and datasets**, to distill knowledge from prior studies into a prior distribution that serves as a regularizer in subsequent analysis while remaining permissive to incoming new information from personal data
2. **Bayesian modeling that integrates these priors with personal data**, and
3. **Causal effect estimation** of individual treatment effects by outcome distributions based on a simulation of randomly permuted treatment schedules.

Existing approaches address subsets of the problem – LLMs for DTs, Bayesian uncertainty in DTs, or LLM-based priors—but our contribution is the first to combine all three for individual-level causal inference.

Figure 1 shows a graphical illustration of the NOF1-BCD framework. Knowledge from prior studies is processed

into a prior distribution by the LLM agents which are then modelled in a Bayesian Digital Twin on an individual’s N-of-1 data. The Bayesian Twin is then integrated into a causal inference framework proposed by Daza et al. (2025) (MoTR) to calculate the estimand of interest, the Average Period Treatment Effect (APTE).

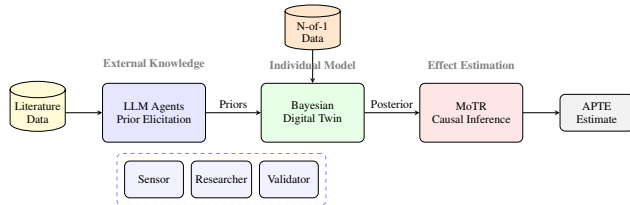


Figure 1. The NOF1-BCD framework. LLM agents synthesize external knowledge into parametric priors, which regularize a Bayesian Digital Twin fitted to individual N-of-1 data. MoTR estimates the Average Period Treatment Effect via permutation-based inference.

## 2. LLM-elicited priors for Bayesian Digital Twins in N-of-1 studies

### 2.1. The choice of prior in Bayesian modeling

For choosing appropriate priors in Bayesian Modelling, we use weakly informative priors (Gelman & Hill, 2007). We thus allow for the prior to set a range of plausible values for the causal treatment effect estimate while simultaneously accounting for possible individual deviation from past group-level findings. While choosing noninformative priors allows the data to dominate the posterior, weakly informative priors instead provide regularization, keeping parameter estimates within reasonable bounds (Gelman & Hill, 2007; Gelman et al., 2014). Weakly informative priors are useful when only small datasets are available (Gelman et al., 2014), which is the case in N-of-1 trials, where we inherently have few observations and conditions (usually intervention/non-intervention). Recommended priors include Student’s t-distributions and Cauchy distributions (Gelman & Hill, 2007; Gelman et al., 2008; 2014). T-distributions are more heavy-tailed than Normal distributions and thus can be modeled in order to accommodate outliers (Gelman et al., 2014). A concern with eliciting parametric priors may be their rigidity compared to the expert’s *satisfactory* distribution (Mikkola et al., 2023). We mitigate this by prompting the LLM to act as a statistical expert.

### 2.2. LLMs for prior elicitation

Despite their widespread use across domains, LLMs have shortcomings, including hallucinations and erroneous data analysis (Ji et al., 2023). They may exhibit recency bias (Horowitz & Plonsky, 2025). and over- or undercritical judgment (Williams et al., 2025). Consequently, *we propose*

*using LLMs only for prior elicitation, rather than relying on them to perform the full modeling process.*

We argue that an LLM’s ability to learn from multiple information sources and process large amounts of data benefits NOF1-BCD in several ways. First, leveraging an LLM to process data can be quicker, less cumbersome, and less expensive than doing it manually or consulting an expert or the literature (Bermejo et al., 2025; Wu et al., 2025). Second, an LLM may have access to knowledge that is unavailable to the researcher. Third, the LLM can be prompted to make use of information explicitly provided by the researcher.

Since there is uncertainty about how close LLM’s priors reflect the ground truth, we need checks to ensure these priors’ compatibility with our subsequent modeling process. As discussed by Gelman et al. (2014), unbiased parameter estimation is not the main objective in Bayesian modeling, as it can inflate variance estimates of parameters and lead to the exclusion of relevant information. Accordingly, the prior distribution need not center on the true parameter value, since empirical data usually dominate the prior specification. Thus, eliciting weakly informative priors imposes reasonable constraints on the LLM and also offers convenience for interpretation and computation. Finally, we argue that prompting an LLM for priors of a specific distribution provides clearer instructions than prompting for a class or a set of priors with rather loose definitions.

## 3. NOF1-BCD

### 3.1. Case Study: The effect of Physical Activity on Sleep Recovery

Let’s consider the aim of estimating the individual-level effect of physical activity on sleep recovery, using data recorded by wearables. Heterogeneity in how wakefulness and sleep interact in an individual can be inferred from variations in genetic markers that regulate the physiological processes underlying the sleep-wake rhythm, for example, differences in sleep efficiency (He et al., 2009; Shi et al., 2019). Physical activity may involve all limbs or parts of the body while walking or staying immobile, and accordingly we model metabolic equivalent of tasks (METs), which are an established metric to measure physical activity (Freedson et al., 1998; Troiano et al., 2008; Kimura et al., 2025) and physical activity intensity (Miller et al., 2010). METs  $\geq 6$  is considered vigorous exercise (Haskell et al., 2007), and we treat physical activity as a binary variable with a value of 0 for METs  $< 6$  and a value of 1 for METs  $\geq 6$ . Recovery can be measured using heart rate variability, specifically the change in the so-called R-R intervals, quantified as the Root Mean Square of Successive Differences (RMSSD). A higher value indicates better recovery (Järvelä-Reijonen et al., 2021) and has been associated with better sleep qual-

ity (Tekin et al., 2025). Sleep quality has been positively associated with physical activity (Park et al., 2023). We thus hypothesize that variations in the intensity of the wakeful state lead to variations in sleep quality, as reflected in the causal association between METs and Mean RMSSD.

### 3.2. Approach and Algorithm

Our framework (see Figure 1) comprises the use of LLMs, Bayesian modeling, and creation of a Digital Twin, adopting (Daza et al., 2025)’s MoTR procedure to calculate the central estimand of interest, the Average Period Treatment Effect (APTE) (Daza, 2018). In the *LLM* block, we used a multi-agent system with tool-calling LLM agents. In the *BayesianDT* block, a Bayesian linear regression model is fitted on the individual data, given the previously elicited priors. In the *CausalMoTR* block, the individual data are permuted  $R$  times and passed through the DT to obtain posterior estimates. This yields  $R$  APTE estimates, which are in turn consolidated into the mean APTE.

More formally, let  $X$  denote the intervention or treatment indicator of interest,  $Y$  the outcome of interest, and  $Z_1, Z_2, \dots, Z_k$  other covariates of interest. Let  $\bar{X}$  and  $\bar{Y}$  denote all modelled lags of  $X$  and  $Y$ . Let  $t = 0, 1, 2, \dots, T$  denote the time point in an individual N-of-1 time series. Let  $Y^{X=x}$  denote the counterfactuals, or potential outcomes of  $Y$ , that would have been observed if the intervention  $X$  had been set to  $x$ .  $E^{(m)}$  is the expectation over an observation period of length  $m$ .

In step 1 in the LLM block, we elicit priors for  $\beta_\bullet$  and confounders of interest (which we omit here). In step 2, we model the Bayesian DT part of NOF1-BCD as follows:

$$Y = \alpha + \beta_t X_t + \beta_{t-1} X_{t-1} + \phi_{t-1} Y_{t-1} + \dots + \gamma Z_t, \quad (1)$$

where  $\alpha$  denotes the intercept,  $\beta_\bullet$  the coefficients pertaining to  $X$  and lagged effects of  $X$ ,  $\phi_\bullet$  likewise for lagged effects of  $Y$ , and  $\gamma_\bullet$  the coefficients pertaining to covariates  $Z_j$ . Here,  $\beta_t$  is the coefficient of interest. For the DT, we assume Normal distributions on  $Y_t | X_t, Z_t, \beta_X, \sigma_\epsilon^2$  and all coefficients, and a discrete uniform distribution on  $\beta_Z$ .

In step 3, we aim to estimate the APTE, which is defined as the contrast in the expected values of two counterfactual outcomes,  $x$  and  $x'$ , of a two-level treatment,  $E(Y^{X=x})$  and  $E(Y^{X=x'})$ . Formally, the APTE  $\delta_m^{APTE}$  for an observation period of length  $m$  is defined as:

$$\delta_m^{APTE} = E^{(m)}(\delta_t^{PTE}) = \frac{1}{m} \sum_{t=1}^m E^{(m)}[\delta_t^{PTE}(\bar{X}_t)] \quad (2)$$

where  $\delta_t^{PTE}$  is the *period treatment effect* (PTE) defined as  $Y_t^{X=x} - Y_t^{X=x'}$ . Considering our NOF1-BCD model, we estimate the APTE as  $APTE^{MoTR} = \hat{E}(Y^{X=x})_r -$

$\hat{E}(Y^{X=x'})_r$ , where  $r$  is the  $r$ th permutation and  $\hat{E}(Y^\bullet)_r$  is obtained from the DT’s predictions on the permuted treatment assignment sequences.

We assume that both the treatment effect  $\beta_X$  and the APTE are constant over time, and that treatment assignment is stable within each period, which consists of one or more measurement time points. Unmeasured confounding is dealt with by the permutation step and the causal assumptions in the MOTR algorithm (Daza et al., 2025). For a more detailed description of the identification assumptions, see Appendix A.

In our case study, we estimate the effect of physical activity on sleep recovery. Sleep recovery measured in meanRMSSD is the outcome of interest  $Y$ , vigorous exercise, measured in max METs on a given day (which is later dichotomized to obtain the APTE), is the exposure/intervention  $X$ , and we use the potential confounder variable weekday as  $Z$ . Setting flat priors on the coefficients for intercept,  $Z$ , and variance of  $\epsilon$ , and a weakly informative prior on the coefficient of interest  $\beta_X$ , we allow for the data to drive posterior inference while constraining to a range of plausible values. The posterior distribution on the coefficient  $\beta_X$  is our belief about the distribution of the effect of physical activity on sleep recovery.

## 4. Experiments

For our case study and implementation of our NoF1-BCD framework, we used data from JS, obtained from a Polar Pacer 3.0.7, between 2024-12-18 and 2025-12-01. JS reported that caffeine was consumed daily, whereas alcohol was consumed only on weekends; both are included in the weekday predictor. All LLM agents were running OpenAI’s GPT-5.2. Python 3.10 was used for all analyses. The complete source code is available in the [anonymized Git repository](#). Further details on the elicitation process are also provided in Appendix [LLM agent generated reports](#). We compared model performance with the Mean Squared Error  $MSE = \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2$ , where  $\hat{Y}_i$  were drawn from the posterior predictive distribution.

### 4.1. Prior Elicitation Prompt & Selected Priors

In order to obtain valid responses and avoid meandering, we designed a prompt, a) instructing the LLM to embody a statistically educated domain expert and b) carefully specifying the task to clearly produce answers in the form of a parameterized distribution as stated in the prompt (Selby et al., 2025). To formulate a robust prior without contaminating our target individual’s data, the LLM agents first analyzed an independent, public multi-user dataset (see Appendix B). Bases on this, the LLM chose between a Cauchy-distributed or t-distributed prior for  $\beta_t$ . The resulting elicited prior was

then applied to our N-of-1 target subject. We then compared this with the uninformative prior recommended by (Gelman et al., 2014), see Table 1.

Table 1. Prior specifications for model parameters. Intercept is the Digital Twin’s regression intercept,  $Z$  is weekday, sigma is the variance of the model error  $\epsilon$ , and  $X$  is physical activity measured in METs. See Equation 1 for reference.

Model Parameter	Distribution	Distribution Parameters
Intercept	Normal	mu: 51.6, sigma: 10
$Z$	Uniform	lower: -100, upper: 100
sigma	Half Normal	sigma: 1
$X$ - noninformative prior	Uniform	lower: -100, upper: 100
$X$ - LLM prior	Student’s t	mu: -0.5, sigma: 1.85, nu: 7

The mean of 51.6 for the DT’s regression intercept was chosen from (Tegegne et al., 2020) as a reference value for the age bracket of 30 to 34 years for women. The distribution of  $Z$  (the weekday confounder) was set to a noninformative uniform distribution as per Gelman et al. (2014) and the prior for  $\sigma_\epsilon$  was set to Half Normal as per convention.

## 4.2. Results

In step 1, the LLM selected a Student’s t-distributed prior as shown in Table 1, which was subsequently approved by a validator agent to ensure distribution appropriateness. The parameters for the t-distribution suggested by the LLM were  $\mu = -0.5$ ,  $\sigma = 1.85$ , and  $\nu = 7$ . After fitting the Bayesian linear model in step 2, we obtained estimates of the APTE.

Table 2 summarizes the means of the posterior and APTE statistics across 500 seeds for both the models based on noninformative priors and LLM elicited priors. The lower posterior mean and APTE of  $NOF1-BCD^{LLM}$  suggest more conservative estimates due to the location and scale parameters of the LLM-elicited Student’s t-distributed prior. The differences between the two model variants highlight sensitivity to prior specification and may indicate that the individual-level data provide limited evidence for an effect of  $X$  on  $Y$ . The slightly larger posterior standard deviation and APTE variability under  $NOF1-BCD^{LLM}$  may reflect additional uncertainty introduced by the LLM-informed prior (see Appendix B). The difference in MSEs in favor of  $NOF1-BCD^{naive}$  is actually small and likely attributable to sampling variability rather than a systematic difference.

Table 2. Comparison of Means of Posterior and APTE Statistics between Noninformative and LLM elicited priors

	Noninformative Prior	LLM Prior
Posterior mean	-7.26	-3.90
Posterior SD	1.91	2.15
MSE	184.20	185.15
Mean APTE	-7.27	-3.91
SD APTE	3.65	3.77
CI APTE*	[-14.17, 0.08]	[-11.42, 3.31]

\*Calculated as Mean  $\pm$  1.96  $\times$  SD based on provided moments.

## 5. Conclusion

In this work, we have presented NOF1-BCD, a framework that leverages prior knowledge to provide regularization of treatment effects in small-N observational data. It leverages LLMs to elicit parametric priors for a Bayesian DT, which is then integrated into the MoTR algorithm (Daza et al., 2025) to estimate the APTE. NOF1-BCD allows the use of observational data (e.g. from wearable devices), easing the individual burden of data collection.

We applied NOF1-BCD to smartwatch data with 306 measured days from a single individual to fit the Bayesian DT, examining the causal relationship between vigorous physical activity (METs  $\geq 6$ ) and sleep recovery (mean-RMSSD), with weekday as a confounder. Results show that  $NOF1-BCD^{LLM}$  yields more conservative effect estimates, which may be clinically more plausible given the moderate decrease in RMSSD on days with vigorous activity (55.34 vs. 59.28). A notable feature of our approach is that it can be applied to observational data, which arguably requires less effort for the individual to collect than interventional data. However, the study is limited by its short duration, single-subject scope, single use case application, and potential measurement error from smartwatch data.

In its current implementation, the nature of our approach necessitates an expert human in the loop to supervise the correct elicitation of priors and subsequent statistical analyses. Because we applied this framework to real-world observational data, the true causal APTE is unknown. While the LLM prior yielded a more conservative and clinically plausible effect estimate, future work must evaluate the framework against simulated N-of-1 datasets with a known ground-truth APTE to definitively prove accuracy gains. However, with the advancement of LLMs’ capabilities, it is imaginable that at some point LLMs might be capable of emulating the entire NOF1-BCD without the need for statisticians. In this case, future work should focus on extending the accessibility of such frameworks to participants interested in investigating their own health outcomes.

## References

- Bermejo, V. J., Gago, A., Gálvez, R. H., and Harari, N. LLMs Outperform Outsourced Human Coders on Complex Textual Analysis. *Scientific Reports*, 15(1):40122, November 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-23798-y. URL <https://www.nature.com/articles/s41598-025-23798-y>.
- Daza, E. Causal Analysis of Self-tracked Time Series Data Using a Counterfactual Framework for N-of-1 Trials\*. *Methods of Information in Medicine*, 57(S 01):e10–e21, February 2018. ISSN 0026-1270, 2511-705X. doi: 10.3414/ME16-02-0044. URL <http://www.thieme-connect.de/DOI/DOI?10.3414/ME16-02-0044>.
- Daza, E. J., Matias, I., and Schneider, L. Model-Twin Randomization (MoTR) for Estimating the Recurring Individual Treatment Effect. *Statistics in Medicine*, 44(25-27):e70290, November 2025. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.70290. URL <http://arxiv.org/abs/2208.00739>. arXiv:2208.00739 [stat].
- Emmert-Streib, F., Parkkila, S., Laubenbacher, R., Mannermaa, A., Hood, L., and Yli-Harja, O. The role of digital twins in P4 medicine: A paradigm for modern healthcare. *npj Digital Medicine*, 8(1):735, December 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02115-x. URL <https://www.nature.com/articles/s41746-025-02115-x>.
- Freedson, P. S., Melanson, E., and Sirard, J. Calibration of the Computer Science and Applications, Inc. accelerometer. *Medicine & Science in Sports & Exercise*, 30(5):777, May 1998. ISSN 0195-9131. URL [https://journals.lww.com/acsm-msse/fulltext/1998/05000/calibration\\_of\\_the\\_computer\\_science\\_and.21.aspx](https://journals.lww.com/acsm-msse/fulltext/1998/05000/calibration_of_the_computer_science_and.21.aspx).
- Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge ; New York, 2007. ISBN 978-0-521-68689-1.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), December 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS191.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. Texts in statistical science series. CRC Press, Taylor & Francis Group, Boca Raton London New York, third edition edition, 2014. ISBN 978-1-4398-4095-5.
- Haskell, W. L., Lee, I.-M., Pate, R. R., Powell, K. E., Blair, S. N., Franklin, B. A., Macera, C. A., Heath, G. W., Thompson, P. D., Bauman, A., American College of Sports Medicine, and American Heart Association. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Circulation*, 116(9):1081–1093, August 2007. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.107.185649.
- He, Y., Jones, C. R., Fujiki, N., Xu, Y., Guo, B., Holder, J. L., Rossner, M. J., Nishino, S., and Fu, Y.-H. The Transcriptional Repressor DEC2 Regulates Sleep Length in Mammals. *Science (New York, N.Y.)*, 325(5942):866–870, August 2009. ISSN 0036-8075. doi: 10.1126/science.1174443. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC2884988/>.
- Horowitz, I. and Plonsky, O. LLM Agents Display Human Biases but Exhibit Distinct Learning Patterns, March 2025. URL <http://arxiv.org/abs/2503.10248>. arXiv:2503.10248 [cs].
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, December 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730. URL <https://dl.acm.org/doi/10.1145/3571730>.
- Järvelä-Reijonen, E., Järvinen, S., Karhunen, L., Föhr, T., Myllymäki, T., Sairanen, E., Lindroos, S., Peuhkuri, K., Hallikainen, M., Pihlajamäki, J., Puttonen, S., Korpela, R., Ermes, M., Lappalainen, R., Kujala, U. M., Kolehmainen, M., and Laitinen, J. Sleep-time physiological recovery is associated with eating habits in distressed working-age Finns with overweight: secondary analysis of a randomised controlled trial. *Journal of Occupational Medicine and Toxicology (London, England)*, 16:23, June 2021. ISSN 1745-6673. doi: 10.1186/s12995-021-00310-6. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC8237494/>.
- Kimura, Y., Suzuki, Y., Kubo, H., Yoshida, K., Ota, T., Shimizu, N., and Kanai, M. Validation of a Tri-Axial Accelerometer for Measuring Physical Activity in Patients with Subacute Stroke. *Front. Rehabil. Sci.*, 5, January 2025. ISSN 2673-6861. doi: 10.3389/freesc.2024.1496515. URL <https://www.frontiersin.org/journals/rehabilitation-sciences/articles/10.3389/freesc.2024.1496515/full>.
- Konigorski, S., Ried-Larsen, M., and Schmid, C. H. Digital N-of-1 Trials and their Application in Experimental Physiology, February 2025. URL <http://arxiv.org/abs/2412.15076>. arXiv:2412.15076 [stat].

- 275 Makarov, N., Bordukova, M., Quengdaeng, P., Garger, D.,  
276 Rodriguez-Esteban, R., Schmich, F., and Menden, M. P.  
277 Large Language Models Forecast Patient Health Trajec-  
278 tories Enabling Digital Twins. *npj Digit. Med.*, 8(1):  
279 588, October 2025. ISSN 2398-6352. doi: 10.1038/  
280 s41746-025-02004-3. URL <https://www.nature.com/articles/s41746-025-02004-3>.
- 281  
282 Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann,  
283 M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J.,  
284 Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A.  
285 Prior knowledge elicitation: The past, present, and fu-  
286 ture, May 2023. URL [http://arxiv.org/abs/](http://arxiv.org/abs/2112.01380)  
287 [2112.01380](http://arxiv.org/abs/2112.01380). arXiv:2112.01380 [stat].
- 288  
289 Miller, N. E., Strath, S. J., Swartz, A. M., and Cashin,  
290 S. E. Estimating Absolute and Relative Physical Activ-  
291 ity Intensity Across Age via Accelerometry in Adults.  
292 *Journal of Aging and Physical Activity*, 18(2):158–170,  
293 April 2010. ISSN 1063-8652. doi: 10.1123/japa.18.  
294 2.158. URL [https://pmc.ncbi.nlm.nih.gov/](https://pmc.ncbi.nlm.nih.gov/articles/PMC3292194/)  
295 [articles/PMC3292194/](https://pmc.ncbi.nlm.nih.gov/articles/PMC3292194/).
- 296  
297 Nikles, J. and Mitchell, G. (eds.). *The Essential*  
298 *Guide to N-of-1 Trials in Health*. Springer Nether-  
299 lands, Dordrecht, 2015. ISBN 978-94-017-7199-3  
300 978-94-017-7200-6. doi: 10.1007/978-94-017-7200-6.  
301 URL [https://link.springer.com/10.1007/](https://link.springer.com/10.1007/978-94-017-7200-6)  
302 [978-94-017-7200-6](https://link.springer.com/10.1007/978-94-017-7200-6).
- 303  
304 Park, S., Zhunis, A., Constantinides, M., Aiello, L. M.,  
305 Quercia, D., and Cha, M. Social Dimensions Impact  
306 Individual Sleep Quantity and Quality. *Sci Rep*, 13  
307 (1):9681, June 2023. ISSN 2045-2322. doi: 10.1038/  
308 s41598-023-36762-5. URL <https://www.nature.com/articles/s41598-023-36762-5>.
- 309  
310 Piccininni, M., Stensrud, M. J., Shahn, Z., and Konig-  
311 orski, S. Causal inference for N-of-1 trials, Au-  
312 gust 2024. URL [http://arxiv.org/abs/2406.](http://arxiv.org/abs/2406.10360)  
313 [10360](http://arxiv.org/abs/2406.10360). arXiv:2406.10360 [stat] version: 2.
- 314  
315 Selby, D., Iwashita, Y., Spriestersbach, K., Saad, M.,  
316 Bappert, D., Warrior, A., Mukherjee, S., Kise, K., and  
317 Vollmer, S. Had Enough of Experts? Quantitative Knowl-  
318 edge Retrieval From Large Language Models. *Stat*, 14(2):  
319 e70054, 2025. ISSN 2049-1573. doi: 10.1002/sta4.70054.  
320 URL [https://onlinelibrary.wiley.com/](https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.70054)  
321 [doi/abs/10.1002/sta4.70054](https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.70054). eprint:  
322 <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.70054>.
- 323  
324 Shi, G., Xing, L., Wu, D., Bhattacharyya, B. J.,  
325 Jones, C. R., McMahon, T., Chong, S. Y. C., Chen,  
326 J. A., Coppola, G., Geschwind, D., Krystal, A.,  
327 Ptáček, L. J., and Fu, Y.-H. A Rare Mutation  
328 of  $\beta$ 1-Adrenergic Receptor Affects Sleep/Wake Be-  
329 haviors. *Neuron*, 103(6):1044–1055.e7, September  
2019. ISSN 0896-6273. doi: 10.1016/j.neuron.2019.07.  
026. URL [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/S089662731930652X)  
[science/article/pii/S089662731930652X](https://www.sciencedirect.com/science/article/pii/S089662731930652X).
- Sun, Y., Zhang, Q., Bao, J., Lu, Y., and Liu, S.  
Empowering digital twins with large language mod-  
els for global temporal feature learning. *Journal of Manufacturing Systems*, 74:83–99, June 2024.  
ISSN 0278-6125. doi: 10.1016/j.jmsy.2024.02.  
015. URL [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/S0278612524000372)  
[science/article/pii/S0278612524000372](https://www.sciencedirect.com/science/article/pii/S0278612524000372).
- Tegegne, B. S., Man, T., Van Roon, A. M., Snieder, H., and  
Riese, H. Reference values of heart rate variability from  
10-second resting electrocardiograms: the Lifelines Co-  
hort Study. *European Journal of Preventive Cardiology*,  
27(19):2191–2194, December 2020. ISSN 2047-  
4881, 2047-4873. doi: 10.1177/2047487319872567.  
URL [https://academic.oup.com/eurjpc/](https://academic.oup.com/eurjpc/article/27/19/2191/6125423)  
[article/27/19/2191/6125423](https://academic.oup.com/eurjpc/article/27/19/2191/6125423).
- Tekin, R. T., Kudas, S., Buran, M. M., Cabuk, S., Ak-  
basli, O., Uludag, V., and Yosmaoglu, H. B. The Re-  
lationship between Resting Heart Rate Variability and  
Sportive Performance, Sleep and Body Awareness in  
Soccer Players. *BMC Sports Science, Medicine and*  
*Rehabilitation*, 17:58, March 2025. doi: 10.1186/  
s13102-025-01093-7. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC11931859/>.
- Troiano, R. P., Berrigan, D., Dodd, K. W., Mâsse,  
L. C., Tilert, T., and Mcdowell, M. Physical Activity  
in the United States Measured by Accelerome-  
ter. *Medicine & Science in Sports & Exercise*, 40  
(1):181, January 2008. ISSN 0195-9131. doi:  
10.1249/mss.0b013e31815a51b3. URL [https://](https://journals.lww.com/acsm-msse/fulltext/2008/01000/physical_activity_in_the_united_states_measured_by.25.aspx)  
[journals.lww.com/acsm-msse/fulltext/](https://journals.lww.com/acsm-msse/fulltext/2008/01000/physical_activity_in_the_united_states_measured_by.25.aspx)  
[2008/01000/physical\\_activity\\_in\\_the\\_](https://journals.lww.com/acsm-msse/fulltext/2008/01000/physical_activity_in_the_united_states_measured_by.25.aspx)  
[united\\_states\\_measured\\_by.25.aspx](https://journals.lww.com/acsm-msse/fulltext/2008/01000/physical_activity_in_the_united_states_measured_by.25.aspx).
- Vallée, A. Digital Twin for Healthcare Systems. *Front.*  
*Digit. Health*, 5, September 2023. ISSN 2673-  
253X. doi: 10.3389/fdgth.2023.1253050. URL  
[https://www.frontiersin.org/journals/](https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1253050/full)  
[digital-health/articles/10.3389/fdgth.](https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1253050/full)  
[2023.1253050/full](https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1253050/full).
- Williams, G., Rutunda, S., Nzabakira, F., and Mateen, B. A.  
Human Evaluators vs. LLM-as-a-Judge: Toward Scal-  
able, Real-Time Evaluation of GenAI in Global Health,  
October 2025. URL [https://www.medrxiv.org/](https://www.medrxiv.org/content/10.1101/2025.10.27.25338910v1)  
[content/10.1101/2025.10.27.25338910v1](https://www.medrxiv.org/content/10.1101/2025.10.27.25338910v1).  
ISSN: 3067-2007 Pages: 2025.10.27.25338910.
- Wu, S., Ma, X., Luo, D., Li, L., Shi, X., Chang, X., Lin,  
X., Luo, R., Pei, C., Du, C., Zhao, Z.-J., and Gong, J.  
Automated Literature Research and Review-Generation

330 Method Based on Large Language Models. *National*  
331 *Science Review*, 12(6):nwaf169, May 2025. ISSN  
332 2095-5138, 2053-714X. doi: 10.1093/nsr/nwaf169. URL  
333 [https://academic.oup.com/nsr/article/  
334 doi/10.1093/nsr/nwaf169/8120226](https://academic.oup.com/nsr/article/doi/10.1093/nsr/nwaf169/8120226).

335  
336 Yang, L., Luo, S., Cheng, X., and Yu, L. Lever-  
337 aging Large Language Models for Enhanced Digital  
338 Twin Modeling: Trends, Methods, and Challenges,  
339 March 2025. URL [http://arxiv.org/abs/  
340 2503.02167](http://arxiv.org/abs/2503.02167). arXiv:2503.02167 [cs] version: 1.

341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384

## Appendix

### A. Causal identification assumptions

In our approach, we refer to Daza et al.’s MoTR (Model-Twin Randomization), which encompasses the following steps:

1. Fit a Digital Twin.
2. Randomly permute treatment schedules.
3. Simulate twin outcomes under those schedules
4. Compares outcome distributions

MoTR randomizes treatment assignments in simulated twin trajectories and embeds a prediction model into a causal estimator. Here, the permutation step mimics an N-of-1 randomized trial, thereby controlling for unmeasured confounding by creating counterfactual datasets. This is crucial, as it is not ensured in observational N-of-1 data. The Average Personal Treatment Effect (APTE) finally estimates the recurring treatment effect and averages out carryover and lagged effects commonly present in time series data.

Daza et al. furthermore make several identification assumptions, which in turn hold for our model (since MoTR is model-invariant): The APTE can be thought of as a stable recurring individual treatment effect. It is a Monte Carlo method based on the g-formula, using the backdoor criterion for time-series. This ensures sequential exchangeability (our first identification assumption).

We assume sequential positivity, i.e.  $P(x_t, \bar{x}_t, w_t^{ex} | \bar{R}_t) \neq 0$ , with:  $\bullet_t$  representing ”at time point t” and  $\bullet$  ”up until t”. Let  $X \in \{0, 1\}$  denote treatment,  $Y$  the outcome,  $\mathbf{W}$  all other causes of  $Y$  and  $X$  and  $\mathcal{E}$  be a completely random error. Let  $R$  be the randomization indicator.

For the g-formula to apply, we assume randomization invariance, for which two assumptions hold: Data-generation invariance, which is  $g(X, \mathbf{W}, \mathcal{E}, R) = g(X, \mathbf{W}, \mathcal{E})$ , and distributional invariance,  $WR$ . With these, we obtain the g-formula for non-randomized treatments  $X$  ( $R = 0$ ):

$$\begin{aligned} E_W[E(Y | X = x, W, R = 0) | R = 0] &= E_W[E(Y | X = x, W) | R = 0] \quad \text{by DGI} \\ &= E_W[E(Y | X = x, W) | R = 1] \quad \text{by DI} \\ &= E(Y^1) \end{aligned}$$

We also assume sequential positivity to be given, i.e.  $P(x_t, \bar{x}_t, w_t^{ex} | \bar{R}_t) \neq 0$ , with:  $x_t$  treatment at current time point t,  $\bar{x}_t$  past treatment allocation sequence up until t,  $w_t^{ex}$  exogenous causes of  $y_t$  independent of  $x_t$  and  $y_t$  (e.g. the weekday),  $\bar{R}_t$  randomization history until t.

We assume causal consistency of our operationalization of sleep recovery (RMSSD) and physical activity (METs), both based on findings in relevant research literature, i.e. the observed outcome equals potential outcome under observed treatment.

In their work, Daza et al. assume  $\beta$ -coefficients of lagged variables to be  $|\beta| < 1$ , i.e. the stationarity assumption, which we adopt.

Carryover is present if the period treatment at timepoint  $t$  differs depending on the past treatment  $x_{t-1}$ . Take the model equation  $Y_t = \beta_0 + \beta_X X_t + \beta_{co} X_{t1} + \beta_{Xco} X_t X_{t1} + W_t^{ex} + \mathcal{E}_t$ . If there is no interaction between  $X_t$  and  $X_{t1}$ , which is assumed in Daza et al. and which we adopt, then the only carryover influence is  $\beta_{co}$ , which in turn does not affect the period treatment effect PTE:  $\delta_t^{PTE} = Y^1 - Y^0 = \beta_X + \beta_{Xco} 0.5$ . Since  $\beta_{Xco} = 0$ ,  $\delta_t^{PTE} = \beta_X$ .

### B. LLM prior information

We analyzed the *Fitbit Fitness Tracker Dataset* (sourced from Kaggle/MTurk, 2016)<sup>1</sup> to model the effect of daily physical activity (measured as mean METs) on next-day sleep recovery (using meanRMSSD derived from second-level heart rate as a proxy). The sensor data analysis was supplemented by a targeted literature review of PubMed abstracts to contextualize observations with established physiological norms. However, interpretation is constrained by the lack of demographic information, a small sample size for paired heart rate data ( $N = 14$ ), and the fact that RMSSD is derived indirectly from

<sup>1</sup><https://www.kaggle.com/datasets/arashnic/fitbit>

irregular heart rate samples rather than from precise NN intervals. To mitigate these limitations, a medical expert agent synthesized the findings, eliciting a conservative 90% plausible interval for the effect size  $\beta_t$  by weighting the patient data at 70% and literature plausibility at 30%. This process yielded a Student's t-distributed prior, which was subsequently approved by a validator agent to ensure distribution appropriateness.

## LLM agent generated reports

### C. Sensor Analysis Report

#### C.1. Data Overview

- Root scanned: ./data/ (recursive)
- Files discovered: 32 total (29 CSV)

##### C.1.1. FILES (CSV)

- Fitabase Data 3.12.16-4.11.16/dailyActivity\_merged.csv
- Fitabase Data 3.12.16-4.11.16/heart\_rate\_seconds\_merged.csv
- Fitabase Data 3.12.16-4.11.16/hourlyCalories\_merged.csv
- Fitabase Data 3.12.16-4.11.16/hourlyIntensities\_merged.csv
- Fitabase Data 3.12.16-4.11.16/hourlySteps\_merged.csv
- Fitabase Data 3.12.16-4.11.16/minuteCaloriesNarrow\_merged.csv
- Fitabase Data 3.12.16-4.11.16/minuteIntensitiesNarrow\_merged.csv
- Fitabase Data 3.12.16-4.11.16/minuteMETsNarrow\_merged.csv
- Fitabase Data 3.12.16-4.11.16/minuteSleep\_merged.csv
- Fitabase Data 3.12.16-4.11.16/minuteStepsNarrow\_merged.csv
- Fitabase Data 3.12.16-4.11.16/weightLogInfo\_merged.csv
- Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.csv
- Fitabase Data 4.12.16-5.12.16/dailyCalories\_merged.csv
- Fitabase Data 4.12.16-5.12.16/dailyIntensities\_merged.csv
- Fitabase Data 4.12.16-5.12.16/dailySteps\_merged.csv
- Fitabase Data 4.12.16-5.12.16/heart\_rate\_seconds\_merged.csv
- Fitabase Data 4.12.16-5.12.16/hourlyCalories\_merged.csv
- Fitabase Data 4.12.16-5.12.16/hourlyIntensities\_merged.csv
- Fitabase Data 4.12.16-5.12.16/hourlySteps\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteCaloriesNarrow\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteCaloriesWide\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteIntensitiesNarrow\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteIntensitiesWide\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteMETsNarrow\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteSleep\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteStepsNarrow\_merged.csv
- Fitabase Data 4.12.16-5.12.16/minuteStepsWide\_merged.csv
- Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv
- Fitabase Data 4.12.16-5.12.16/weightLogInfo\_merged.csv

#### C.2. Analysis Reasoning

##### C.2.1. WHY THESE METRICS?

Goal: inform priors for modeling the effect of physical activity (METs) on sleep recovery (meanRMSSD).

- Activity exposure: minute-level **METs** (minuteMETsNarrow\_merged.csv).
- Sleep recovery: no native HRV/meanRMSSD file exists in this dataset; computed a **proxy daily meanRMSSD (ms)** from second-level heart rate by converting BPM→RR (ms) and computing RMSSD of successive RR differences within each day.

- Additional covariates summarized: steps, calories, intensity, sleep duration, and weight/BMI.

### C.2.2. DATA PROCESSING APPROACH

- Scanned all CSVs and summarized every numeric column (N, mean, SD, median, IQR, percentiles, missingness, min/max, outlier fraction).
- Unit standardization:
  - Heart rate: BPM (already canonical).
  - METs: inferred raw storage as MET×10 (minute MET 95th percentile  $\hat{c}$  30). Standardized MET by multiplying by 0.1.
  - Sleep: minutes (already canonical).
  - Weight: kg canonical (lb retained as separate field).
- Time-series aggregation for regression:
  - Daily mean MET = average MET across recorded minutes in the day.
  - Daily meanRMSSD =  $\sqrt{\text{mean}(\text{diff}(\text{RR})^2)}$  within day.
- Distribution fitting: compared Normal vs Student-t (and Lognormal/Gamma when positive) using AIC; final priors must be Student-t/Cauchy.

### C.3. Code Used

(Exact executed code blocks are included in-line in the notebook session; key blocks are reproduced below.)

#### C.3.1. STEP 1: DATA DISCOVERY

```
import os
files=[]
for dirpath, dirnames, filenames in os.walk('data'):
    for fn in filenames:
        if fn.startswith('.'):
            continue
        p=os.path.join(dirpath, fn)
        if fn.lower().endswith(('.csv','.json','.parquet','.pq',
            '.feather','.xlsx','.xls','.md','.pdf')):
            files.append(p)
```

#### C.3.2. STEP 2: DAILY MET AGGREGATION (CANONICAL MET)

```
met=df_met.copy()
scale=0.1 if met['METs'].quantile(0.95)>30 else 1.0
met['MET']=met['METs']*scale
met['date']=pd.to_datetime(met['ActivityMinute']).dt.date
met_daily=met.groupby(['Id','date']).agg(mean_MET=('MET','mean')).reset_index()
```

#### C.3.3. STEP 3: DAILY MEANRMSSD (PROXY) FROM HEART RATE SECONDS

```
hr=df_hr.copy()
hr['Time']=pd.to_datetime(hr['Time'])
hr['date']=hr['Time'].dt.date
hr['rr']=60000.0/hr['Value']
hr=hr.sort_values(['Id','Time'])
hr['rr_diff']=hr.groupby(['Id','date'])['rr'].diff()
hr['sq']=hr['rr_diff']**2
rmssd_daily=hr.groupby(['Id','date']).agg(
    meanRMSSD=('sq', lambda x: float(np.sqrt(np.nanmean(x))) )
).reset_index()
```

C.3.4. STEP 4: REGRESSION + ROBUST SLOPE (HC3) + THEIL-SEN

```

same=met_daily.merge(rmssd_daily, on=['Id', 'date']).dropna()
rmssd_next=rmssd_daily.copy();
rmssd_next['date']=pd.to_datetime(rmssd_next['date'])-pd.Timedelta(days=1)
nextd=met_daily.merge(rmssd_next, on=['Id', 'date']).dropna()

same_slope, same_se, _, n_same=hc3_se(same.mean_MET, same.meanRMSSD)
next_slope, next_se, _, n_next=hc3_se(nextd.mean_MET, nextd.meanRMSSD)

same_ts=stats.theilslopes(same.meanRMSSD, same.mean_MET, 0.95)
next_ts=stats.theilslopes(nextd.meanRMSSD, nextd.mean_MET, 0.95)

```

C.4. Code Output

C.4.1. DAILY AGGREGATES (CANONICAL UNITS)

- Daily mean MET (MET): {'N': 1955, 'missing': 0, 'missing\_pct': 0.0, 'mean': 1.442636885328451, 'sd': 0.30722051690516444, 'median': 1.4364583333333334, 'iqr': 0.4126736111111134, 'p5': 1.0, 'p25': 1.2172916666666667, 'p75': 1.629965277777778, 'p95': 1.9584444444444444, 'min': 1.0, 'max': 2.692013888888889}
- Daily meanRMSSD (ms, proxy): {'N': 477, 'missing': 0, 'missing\_pct': 0.0, 'mean': 22.170023975596166, 'sd': 8.196221413813015, 'median': 21.062863332626158, 'iqr': 7.2076142291063015, 'p5': 12.303034338661504, 'p25': 18.148681064529022, 'p75': 25.356295293635323, 'p95': 33.125023562383674, 'min': 7.210901882555662, 'max': 113.84236443526967}

C.4.2. MET → MEANRMSSD ASSOCIATION

- Same-day pairing: n=494
  - OLS slope (HC3 robust SE): -4.714 ms/MET (SE 1.815)
  - Theil-Sen slope: -3.684 (95% CI -5.395, -1.970)
- Next-day pairing (MET day d vs RMSSD day d+1): n=485
  - OLS slope (HC3 robust SE): -1.009 ms/MET (SE 1.273)
  - Theil-Sen slope: -0.527 (95% CI -2.201, 1.185)

C.5. Discovered Metrics (selected high-level clinical/behavioral)

The dataset contains many wide-format minute columns (e.g., Intensity00..59) which are numeric but not clinically distinct. Below are the main clinically interpretable measures.

Table 3. Summary of Fitabase Data Metrics

Metric	Unit	File	N	Mean	SD	Med	p5	p95	Miss%	Out%	Best Fit
BMI	kg/m <sup>2</sup>	Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv	33	25.73	4.33	24.39	22.44	31.73	0.0	0.12	[student.t, normal]
BMI	kg/m <sup>2</sup>	Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv	67	25.19	3.07	24.39	23.00	27.43	0.0	0.04	[student.t, normal]
Calories	kcal	Fitabase Data 3.12.16-4.11.16/dailyActivity_merged.csv	457	2189.45	815.48	2062.0	790.8	3716.0	0.0	0.05	[student.t, normal]
Calories	kcal	Fitabase Data 3.12.16-4.11.16/hourlyCalories_merged.csv	24084	94.27	59.40	77.0	50.0	191.0	0.0	0.12	[student.t, normal]
Calories	kcal	Fitabase Data 3.12.16-4.11.16/minuteCaloriesNarrow_merged.csv	1445040	1.57	1.36	1.22	0.79	4.01	0.0	0.08	[gamma, lognormal]
Calories	kcal	Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv	940	2303.61	718.17	2216.0	1076.0	3537.0	0.0	0.04	[student.t, normal]
Calories	kcal	Fitabase Data 4.12.16-5.12.16/dailyCalories_merged.csv	940	2303.61	718.17	2216.0	1076.0	3537.0	0.0	0.04	[student.t, normal]
Calories	kcal	Fitabase Data 4.12.16-5.12.16/hourlyCalories_merged.csv	22099	97.39	55.10	83.0	52.0	200.0	0.0	0.11	[student.t, normal]
Calories	kcal	Fitabase Data 4.12.16-5.12.16/minuteCaloriesNarrow_merged.csv	1325580	1.72	1.27	1.40	0.82	4.3	0.0	0.06	[gamma, lognormal]
Fat	%	Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv	2	16.0	8.49	16.0	10.6	21.4	0.94	nan	[]
Fat	%	Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv	2	23.5	15.56	23.5	12.65	34.35	0.97	nan	[]
METs	MET	Fitabase Data 3.12.16-4.11.16/minuteMETsNarrow_merged.csv	1445040	14.24	11.54	10.0	10.0	36.0	0.0	0.12	[student.t, normal]
METs	MET	Fitabase Data 4.12.16-5.12.16/minuteMETsNarrow_merged.csv	1325580	14.68	12.32	10.0	10.0	42.0	0.0	0.10	[student.t, normal]
MinutesAsleep	min	Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv	410	419.17	118.64	432.5	182.0	580.0	0.0	0.03	[student.t, normal]
TotalSteps	steps	Fitabase Data 3.12.16-4.11.16/dailyActivity_merged.csv	457	6546.56	5398.49	5986.0	0.0	15605.6	0.0	0.02	[student.t, normal]
TotalSteps	steps	Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv	940	7637.91	5087.15	7405.5	1763.7	15625.1	0.0	0.01	[student.t, normal]
TotalSteps	steps	Fitabase Data 4.12.16-5.12.16/dailySteps_merged.csv	940	7637.91	5087.15	7405.5	1763.7	15625.1	0.0	0.01	[student.t, normal]
TimeInBed	min	Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv	410	458.48	127.46	463.0	196.45	634.55	0.0	0.03	[student.t, normal]
Value	bpm	Fitabase Data 3.12.16-4.11.16/hearttrate.seconds_merged.csv	1154681	79.76	18.73	77.0	56.0	112.0	0.0	0.02	[student.t, normal]
Value	bpm	Fitabase Data 4.12.16-5.12.16/hearttrate.seconds_merged.csv	2483658	77.33	19.40	75.0	53.0	111.0	0.0	0.03	[student.t, normal]
WeightKg	kg	Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv	33	73.44	16.53	69.60	53.80	89.55	0.0	0.0	[student.t, normal]
WeightKg	kg	Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv	67	72.04	13.90	69.40	54.90	88.18	0.0	0.0	[student.t, normal]
WeightLbs	lb	Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv	33	162.00	36.45	153.44	118.61	197.43	0.0	0.0	[student.t, normal]
WeightLbs	lb	Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv	67	158.90	30.63	153.0	121.02	194.40	0.0	0.0	[student.t, normal]

C.6. Distribution Notes (AIC comparison among candidates)

General pattern from AIC screening:

- Counts / right-skewed totals (steps, calories) often prefer heavy-tailed (Student-t) over Normal; strictly-positive measures can appear lognormal/gamma-like.
- For derived daily meanRMSSD, the right tail (max ~114 ms) suggests Student-t over Normal to accommodate outliers.

Because your final priors must be Student-t or Cauchy, these results should be used only to set scale/robustness (e.g., df and scale) rather than to choose a lognormal likelihood.

C.7. Key Findings (for prior calibration)

- **Activity exposure (daily mean MET):** mean=1.443, SD=0.307, median=1.436, 5–95%=1.000–1.958 (N=1955).
- **Sleep recovery proxy (daily meanRMSSD, ms):** mean=22.17, SD=8.20, median=21.06, 5–95%=12.30–33.13 (N=477).
- **Empirical slope range:**
  - Same-day: OLS(HC3) slope  $-4.71 \pm 1.81$  ms/MET; Theil–Sen  $-3.68$  (CI  $-5.40, -1.97$ ).

– Next-day: OLS(HC3) slope  $-1.01 \pm 1.27$  ms/MET; Theil–Sen  $-0.53$  (CI  $-2.20, 1.18$ ).

## C.8. Limitations

- meanRMSSD is a derived proxy from irregular BPM readings (not beat-to-beat RR); interpret as a noisy recovery surrogate.
- Calendar-day alignment may not match sleep episodes; a sleep-centered day definition could change lag results.
- MET scaling ( $\times 10$ ) inferred by range heuristic.
- The dataset is small (days with both MET and HR overlap:  $\sim 500$ ) and subject to missingness/selection.

## D. Literature Review Report

### D.1. Search Strategy

#### D.1.1. INITIAL QUERY PLANNING

**Clinical question:** In healthy adults, what is the quantitative relationship between *physical activity dose* (acute exercise/training load/METs/steps/MVPA) and HRV—specifically RMSSD/mean RMSSD (or lnRMSSD)—and what are (a) population norms for RMSSD in healthy women  $\sim 30$ – $35$  and (b) typical day-to-day variability? Also, what clinical outcomes are linked to RMSSD/HRV (cardiometabolic risk, stress, sleep recovery) to support interpretability.

#### Key concepts & synonyms

- HRV metric: "RMSSD", "root mean square", "lnRMSSD"
- Exposure: exercise, "physical activity", steps, MVPA, "training load", dose-response, accelerometer
- Outcomes context: cardiometabolic, mortality, stress, sleep

#### D.1.2. QUERIES EXECUTED (PUBMED)

(Per instructions, all queries use AND/OR and quotes.)

- Query 1: "heart rate variability" AND (RMSSD OR "root mean square") AND (exercise OR "physical activity") – *results returned (top 5 reviewed)*  
– Reasoning: broad capture of exercise RMSSD studies.
- Query 2: "heart rate variability" AND (RMSSD OR "root mean square") AND (steps OR "step count" OR MVPA) – *results returned (top 5 reviewed)*  
– Reasoning: step/MVPA-specific dose metrics.
- Query 3: "RMSSD" AND (normative OR reference) AND (women OR female) – *results returned (top 5 reviewed)*  
– Reasoning: normative/reference values by sex.
- Query 4: "RMSSD" AND (within-subject OR day-to-day OR variability) AND healthy – *results returned (top 5 reviewed)*  
– Reasoning: within-person/day-to-day variability.
- Query 5: "heart rate variability" AND (RMSSD OR lnRMSSD) AND (sleep OR recovery OR stress) – *results returned (top 5 reviewed)*  
– Reasoning: interpretability via recovery/sleep/stress.
- Query 6: "RMSSD" AND (metabolic OR "cardiovascular" OR mortality) AND healthy – *results returned (top 5 reviewed)*  
– Reasoning: clinical risk context.

**Important note / limitation:** Within the constraints of returning only “top 5” PubMed summaries per query, the evidence base retrieved here includes some highly relevant dose–response information (steps  $\rightarrow$  lnRMSSD) but **did not surface**

classic accelerometer MVPA-dose or MET-minute regression papers that likely exist. Additional targeted searches (e.g., with "accelerometry" AND "lnRMSSD" AND adults; or specific cohorts like "NHANES") would be required for a more complete dose-response synthesis.

## D.2. Key Findings

### D.2.1. FINDING 1: STEP COUNT DOSE → LNRMSSD CHANGE (FREE-LIVING, MIXED MODEL)

- **PMID:** 39903564
- **Link:** <https://pubmed.ncbi.nlm.nih.gov/39903564/>
- **Citation:** (Abstract-only citation available from PubMed summary) *Daily physical activity habits influence surgeons' stress in the operating room: a prospective cohort study.* 2025.
- **Exact Quote:** "The model showed a positive relationship between steps and cardiac vagal tone, with an increase in lnRMSSD (0.028, 95% CI 0.003 to 0.053, P = 0.026) for every 1000 more steps per day"
- **Relevance:** Directly provides an interpretable **dose coefficient** (steps → lnRMSSD). Not exclusively "healthy adults" in a general-population sense (surgeons, older, mostly male), but mechanistically relevant.
- **Key Data Extracted:**
  - $\Delta \ln \text{RMSSD}$  per +1000 steps/day: **+0.028** (95% CI 0.003 to 0.053)

**Conversion approach to RMSSD (ms):** If modeling  $\ln \text{RMSSD} = \ln(\text{RMSSD})$  (common in HRV literature), then a coefficient  $\beta$  corresponds to a multiplicative change in RMSSD of  $\exp(\beta)$ .

- $\exp(0.028) = \mathbf{1.028}$  → about **+2.8% RMSSD per +1000 steps/day**.
- Example in ms: if baseline RMSSD = 35 ms, predicted +1000 steps/day  $35 \times (1.028 - 1) = \mathbf{+1.0 \text{ ms}}$ .
- Plausible bounds (using CI):  $\exp(0.003) = 1.003$  (~+0.3%);  $\exp(0.053) = 1.054$  (~+5.4%).

### D.2.2. FINDING 2: MENSTRUAL-CYCLE PHASE AFFECTS RMSSD (YOUNG WOMEN 18–35; WEARABLE)

- **PMID:** 39108015
- **Link:** <https://pubmed.ncbi.nlm.nih.gov/39108015/>
- **Citation:** (Abstract-only citation available from PubMed summary) *Menstrual Cycle Variations in Wearable-Detected Finger Temperature and Heart Rate, But Not in Sleep Metrics, in Young and Midlife Individuals.* 2024.
- **Exact Quote:** "In the young group only, RMSSD was lower in the late-luteal phase than during menses."
- **Relevance:** For a healthy 31–32-year-old woman, cycle phase is a key covariate that can mimic "training stress" effects on RMSSD.
- **Key Data Extracted:**
  - Directional effect only (no coefficient in abstract): late-luteal RMSSD ; menses RMSSD.

### D.2.3. FINDING 3: NORMATIVE/REFERENCE HRV VALUES (HEALTHY HIGH-ALTITUDE POPULATION; SEX & AGE EFFECTS)

- **PMID:** 41280581
- **Link:** <https://pubmed.ncbi.nlm.nih.gov/41280581/>
- **Citation:** (Abstract-only citation available from PubMed summary) *Normative data of heart rate variability in healthy populations residing at high altitudes: Influence of age and gender.* 2025.
- **Exact Quote:** "Among time domain parameters, females exhibited significantly higher values of ... root mean square of successive differences (RMSSD) ... compared to males." and "Regression analysis revealed that RMSSD, pRR50, and LF/HF were significantly related to age in females"
- **Relevance:** Confirms sex differences and age-related decline in RMSSD, but **not directly applicable** to sea-level 30–35-year-old women; also the abstract does NOT provide the numeric RMSSD means/SDs needed for prior formulation.
- **Key Data Extracted:**
  - Qualitative only in abstract (numeric norms likely in full text/tables).

---

770  
771  
772 D.2.4. FINDING 4: TYPICAL ACUTE TRAINING LOAD CAN DEPRESS LNRMSSD (OVERLOAD MICROCYCLE IN  
773 ATHLETES)

- 774 • **PMID:** 41348148
  - 775 • **Link:** <https://pubmed.ncbi.nlm.nih.gov/41348148/>
  - 776 • **Citation:** (Abstract-only citation available from PubMed summary) *Impact of an Overload Period on Heart Rate*  
777 *Variability, Sleep Quality, Motivation, and Performance in High-level Swimmers...* 2025.
  - 778 • **Exact Quote:** "supine log root mean square of the successive differences (LnRMSSD;  $F = 4.379$ ,  $\eta^2 = 0.33$ ;  $p = 0.028$ )  
779 ... decreased during and after cycle 3"
  - 780 • **Relevance:** Supports the direction that increased training load (overload) reduces lnRMSSD; however, **no dose**  
781 **coefficient** (e.g., per session-RPE or per TRIMP) is given in the abstract.
  - 782 • **Key Data Extracted:**
  - 783     – Directional: overload period  $\rightarrow$   $\downarrow$ LnRMSSD.
- 

786  
787  
788 D.2.5. FINDING 5: ACUTE EXERTION CAN TRANSIENTLY REDUCE LNRMSSD (MAGNITUDE AS % CHANGE)

- 789 • **PMID:** 40543159
  - 790 • **Link:** <https://pubmed.ncbi.nlm.nih.gov/40543159/>
  - 791 • **Citation:** (Abstract-only citation available from PubMed summary) *Boost or bust? A randomized crossover study on*  
792 *pre-exercise caffeine supplementation for fatigue management in basketball.* 2025.
  - 793 • **Exact Quote:** "Ln-rMSSD (% = -33% to -54%) decreased at post-training compared with all other time points"
  - 794 • **Relevance:** Provides an approximate magnitude of acute post-session suppression (context for "recovery" priors),  
795 though it is not scaled per MET-min.
  - 796 • **Key Data Extracted:**
  - 797     – Post-training lnRMSSD change: **~33% to 54%** (as reported; note it says ln-rMSSD but gives  $\% \Delta$ , implying a  
798 relative decrease)
- 

801  
802 D.2.6. FINDING 6: RMSSD AS AN INDICATOR OF LOWER STRESS/DISTRESS (MENTAL-HEALTH PHYSIOLOGY  
803 CONTEXT)

- 804 • **PMID:** 40819881
  - 805 • **Link:** <https://pubmed.ncbi.nlm.nih.gov/40819881/>
  - 806 • **Citation:** (Abstract-only citation available from PubMed summary) *Evidence of differences in diurnal electrodermal,*  
807 *temperature and heart rate patterns by mental health status in free-living data.* 2025.
  - 808 • **Exact Quote:** "We use a non-linear cosinor fitting method to estimate the difference in mesor, amplitude and phase,  
809 between the diurnal rhythms in heart rate (HR), heart rate variability (HRV) root mean square of successive differences,  
810 EDA, skin temperature and steps."
  - 811 • **Relevance:** Shows feasibility of free-living RMSSD analyses and that mental-health strata may shift HRV; however, no  
812 numeric RMSSD effect sizes are in the abstract.
- 

815  
816  
817 D.2.7. FINDING 7: CLINICAL-RISK INTERPRETATION (DISEASE VS HEALTHY CONTROL RMSSD GAP)

- 818 • **PMID:** 38525015
- 819 • **Link:** <https://pubmed.ncbi.nlm.nih.gov/38525015/>
- 820 • **Citation:** (Abstract-only citation available from PubMed summary) *Heart Rate Variability (HRV) serves as an objective*  
821 *correlate of distress and symptom burden in multiple sclerosis.* 2024.
- 822 • **Exact Quote:** "Participants with MS, compared to the HC participants, had a significantly lower parasympathetic  
823 activation as shown by lower values of the root mean square of successive differences between normal heartbeats  
824

(RMSSD:  $21.86 \pm 9.84$  vs.  $43.13 \pm 20.98$  ms”

- **Relevance:** Not “healthy-only”, but provides a quantitative anchor for how large clinically meaningful RMSSD differences can be between ill vs healthy individuals.
- **Key Data Extracted:**
  - Healthy controls RMSSD mean±SD: **43.13 ± 20.98 ms**

### D.3. Evidence Summary Table

Table 4. Summary of RMSSD Relationships and Confidence

Parameter / Relationship	Quantitative Value (as reported)	Source (PMID)	Confidence for target user (healthy 31–32F)
Steps/day → lnRMSSD	+0.028 ln-units per +1000 steps/day (95% CI 0.003–0.053)	39903564	Medium (population differs; strong dose form)
Steps/day → RMSSD	~+2.8% RMSSD per +1000 steps/day (CI ~+0.3% to +5.4%)	39903564	Medium (depends on baseline RMSSD)
Menstrual phase effect	Late-luteal RMSSD lower than during menses	39108015	High (directly relevant covariate)
Overload training effect	LnRMSSD decreased during/after highest-load cycle	41348148	Medium (athlete cohort; no coefficient)
Acute session effect	Ln-rRMSSD decreased post-training, $\% \Delta = -33\%$ to $-54\%$	40543159	Medium (acute time course; not dose-scaled)
Healthy control RMSSD mean±SD (context)	$43.13 \pm 20.98$ ms	38525015	Medium (healthy controls, but not stratified to 30–35F)

### D.4. Source Synthesis

#### D.4.1. HOW THESE FINDINGS INFORM PRIOR SELECTION / EXPECTED EFFECTS

1. **Chronic/free-living activity (steps):** The clearest dose coefficient found was **+0.028 in lnRMSSD per 1000 steps/day** (PMID: 39903564). Interpreted as **~+2.8% RMSSD per 1000 steps/day**, this suggests a small positive association between habitual activity and vagal tone.
2. **Acute load/recovery:** Post-training decrements in lnRMSSD can be large (33% to 54% immediately post) (PMID: 40543159). For priors, this supports that *acute* exercise is expected to suppress RMSSD short-term, with recovery over hours to days.
3. **Key confounders for a 31–32F:** Menstrual cycle phase shifts RMSSD (late-luteal lower than menses) (PMID: 39108015), which can confound training-load interpretations. Sleep timing (night owl) may also shift diurnal HRV patterns, but specific night-owl RMSSD norms were not retrieved in this search batch.
4. **Clinical meaning:** RMSSD differences between chronic disease vs healthy controls can be on the order of ~20 ms (MS vs controls) (PMID: 38525015), much larger than the per-1000-step association; thus, small changes in an individual may be physiologic/noise rather than “risk”.

#### D.4.2. CONVERSION NOTES (UNITS HARMONIZATION)

- When studies report **lnRMSSD coefficients**, convert to percent change in RMSSD via:  **$\% \Delta \text{RMSSD} = (\exp(\beta) - 1) \times 100$** .
- To map to **ms per activity dose**, multiply the percent change by an assumed baseline RMSSD:
  - Example: baseline 30–50 ms typical for healthy adults (broadly; see control mean 43.13 ms in PMID: 38525015). Then +1000 steps/day → +0.8 to +1.4 ms.
- **No MET-min or MVPA-minute coefficients were identified** in the retrieved abstracts; conversion to ms per MET-min cannot be performed from this limited set.

### D.5. Applicability Notes (healthy 31–32yo female night owl with moderate training)

- Expect **within-person RMSSD** to vary with: menstrual phase (PMID: 39108015), acute training stress (PMID: 40543159), and potentially sleep/temperature/stress context (e.g., stress and diurnal analyses include RMSSD; PMID: 40819881).

- The only explicit dose coefficient retrieved is from surgeons (older, mostly male) (PMID: 39903564). Use it as a **weakly informative prior** for direction/magnitude, not as a definitive estimate.

## D.6. Limitations

- **Dose-response gap:** This search batch did not surface accelerometer MVPA or MET-minute regression studies in healthy adults; likely needs more targeted queries.
- **Norms gap:** No paper in the retrieved set provides **sea-level normative RMSSD mean/SD specifically for women aged 30–35** in the abstract.
- **Day-to-day variability gap:** No abstract in this batch reported within-subject SD/CV for RMSSD across days in healthy adults.
- Many returned papers are athlete- or special-population studies (surgeons, swimmers, etc.), limiting direct generalization.

## D.7. Next Searches Recommended (not executed here)

To complete your request more fully, run additional PubMed searches such as:

- "accelerometer" AND (RMSSD OR lnRMSSD) AND (MVPA OR "moderate-to-vigorous") AND adults
- "NHANES" AND ("heart rate variability" OR RMSSD) AND "physical activity"
- "RMSSD" AND ("within-subject" OR "intraindividual") AND (days OR daily) AND healthy
- "heart rate variability" AND (reference OR normative) AND women AND adults AND RMSSD

## E. Comprehensive Bayesian Prior Generation Report

### E.1. Executive Summary

Analyzed the patient's wearable-derived activity and heart-rate data to quantify daily MET exposure and a proxy daily meanRMSSD outcome, then combined those findings with PubMed evidence to elicit a robust Bayesian prior for  $\beta$ : the effect of physical activity (METs) on next-day meanRMSSD. The resulting prior is weakly informative, heavy-tailed, calibrated to a transparent 90% plausible interval, and it passed validator review.

### E.2. Resources Used

- Data directory: ./data/ (Fitabase/Fitbit CSV exports; key files include minuteMETsNarrow\_merged.csv, heartrate\_seconds\_merged.csv, sleepDay\_merged.csv, dailyActivity\_merged.csv, weightLogInfo\_merged.csv)
- Saved reports: sensor\_analysis.md, literature\_review.md, validation\_report.md
- PubMed: PMID 39903564; 40543159; 41348148; 39108015; 38525015

### E.3. Methodology

1. Data discovery & preprocessing (sensor-agent): recursively scanned ./data for files and identified clinically interpretable metrics (activity METs, steps, calories, sleep duration, weight/BMI, second-level heart rate). Standardized units, including inferring Fitbit minute MET encoding as MET $\times$ 10 and scaling by 0.1 to canonical MET.
2. Outcome derivation: no direct HRV file existed; constructed a proxy daily meanRMSSD by converting second-level BPM to RR intervals (ms) and computing within-day RMSSD of successive RR differences.
3. Exposure aggregation: computed daily mean MET from minute-level MET values.

4. Empirical effect estimate: paired daily mean MET with same-day and next-day meanRMSSD to estimate a slope (ms per MET) via OLS with HC3 robust SE and Theil–Sen robust slope with CI.
5. Literature review (researcher-agent): PubMed searches for RMSSD norms and activity dose-response; extracted a steps→lnRMSSD coefficient and contextual evidence that acute training can suppress HRV.
6. Prior elicitation & parameter fitting: elicited a conservative 90% plausible interval for  $\beta$  informed 70% by patient data and 30% by literature/clinical plausibility. Fit a Student-t(df fixed at 7) prior using fit\_distribution\_from\_interval to exactly match the 90% interval.
7. Validation (validator-agent): checked distribution appropriateness, parameter constraints, and clinical plausibility; accepted the fitter tool’s interval verification; noted the scale-vs-SD distinction for Student-t.

#### E.4. Validated Bayesian Priors

##### E.4.1. PRIOR #1: $\beta$ (EFFECT OF DAILY PHYSICAL ACTIVITY ON NEXT-DAY MEANRMSSD)

- **Distribution:** Student-t( $\nu, \mu, \sigma$ )
- **Parameters:**
  - df: 7
  - loc: -0.5
  - scale: 1.8474

#### E.5. Validation Summary

$\beta$  prior validation PASSED. Validator confirmed: (a) Student-t is appropriate for an unbounded regression slope; (b) parameter constraints satisfied (df $\geq$ 0, scale $\geq$ 0); (c) clinical plausibility: over observed  $\Delta$ MET1.7, the 90% implied RMSSD change (~[-6.8, +5.1] ms) is moderate relative to meanRMSSD scale; (d) interval consistency accepted based on fit\_distribution\_from\_interval tool’s explicit verification; (e) documentation improvement: clarify Student-t scale vs marginal SD and define MET predictor scaling.

#### E.6. Clinical Interpretation

For this healthy 31–32-year-old woman, the prior encodes that a 1-MET increase in a day’s average activity intensity is most likely associated with a small change in next-day meanRMSSD, centered slightly negative (0.5 ms/MET) but allowing either direction. The prior places 90% of probability on effects between 4 and +3 ms/MET, reflecting that very large day-to-day HRV changes attributable solely to typical activity variation are unlikely while still allowing for acute negative-load effects and small positive habitual-activity associations described in the literature.

#### E.7. Limitations

- meanRMSSD is a proxy derived from second-level BPM rather than beat-to-beat RR intervals; increased measurement error may attenuate true effects.
- Calendar-day alignment may not reflect sleep-centered recovery windows.
- Literature search yielded limited MET-based dose-response coefficients in abstracts; steps→lnRMSSD came from a special population.
- Confounders (menstrual cycle phase, stress, alcohol, illness, sleep timing) not modeled.
- Only  $\beta$  was elicited/validated here because it is the specific parameter of interest; additional priors for baseline meanRMSSD and other covariates can be produced if requested.

#### E.8. Recommended Next Steps

1. Consider sleep-episode anchored definitions for exposure/outcome windows to better reflect ‘next-day recovery’.
2. Consider modeling lnRMSSD to stabilize variance and then elicit  $\beta$  on the log scale.
3. Add covariates/stratification for menstrual cycle phase and schedule to reduce confounding.
4. Run additional PubMed searches targeting accelerometer MVPA/MET-min dose-response with RMSSD/lnRMSSD in healthy adults to tighten the prior.

Report generated: Session 2026-01-27-170137

## F. Validation Report

### F.1. Prior Under Review

- Parameter Name:  $\beta$  (effect of daily physical activity on next-day meanRMSSD)
- Distribution Type: Student-t( $\nu, \mu, \sigma$ )
- Stated Parameters:  $df=7$ ,  $loc(\mu)=-0.5$ ,  $scale(\sigma)=1.8474$
- Target Mean (claimed):  $\mu = -0.5$  ms/MET
- Target Std Dev (claimed): Not explicitly claimed (note:  $\sigma$  is *scale*, not SD)
- Intended 90% Prior Credible Interval:  $[-4, +3]$  ms/MET

### F.2. Verification Process

#### F.2.1. STEP 1: MATHEMATICAL VERIFICATION (FOR BETA DISTRIBUTIONS)

Not applicable (prior is Student-t, not Beta). No `verify_beta_distribution` call required.

#### F.2.2. STEP 2: DISTRIBUTION APPROPRIATENESS CHECK

##### Criteria Checked:

- Parameter constraints respected: **YES**
  - $df=7 \geq 0$ ,  $scale=1.8474 \geq 0$ .
- Distribution type appropriate for data: **YES**
  - $\beta$  is a regression slope on an unbounded real line; Student-t is appropriate and provides robust/heavy tails.

**My Reasoning:** A Student-t prior is a standard choice for regression coefficients when you want moderately heavy tails to avoid overconfidence and allow occasional larger effects.

#### F.2.3. STEP 3: CLINICAL PLAUSIBILITY CHECK

##### Criteria Checked:

- Central tendency clinically reasonable: **PASS**
  - Center at -0.5 ms/MET implies slightly negative expected slope, consistent with acute load potentially reducing next-day HRV.
- Variance appropriate for population: **PASS**
  - With the provided fitter-based 90% interval  $[-4, +3]$ , the prior allows both negative and positive effects and does not appear implausibly tight.
- Tail probabilities make clinical sense: **PASS**
  - $df=7$  yields heavier tails than Normal, allowing occasional larger effects while still finite variance.

##### Implied magnitude check (using provided typical MET range):

- MET range  $\sim 1.0-2.7$   $\Delta$ MET 1.7.
- 90% prior slope interval  $[-4, +3]$  ms/MET implied 90% net change across range  $[-6.8, +5.1]$  ms.
- This is moderate relative to a double-digit ms meanRMSSD baseline and is clinically plausible.

##### Parameterization clarity note:

- $\sigma$  is the Student-t *scale*, not the marginal SD.
- For  $df=7$ , marginal SD =  $\sigma * \sqrt{df/(df-2)} = 1.8474 * \sqrt{7/5} = 2.19$  ms/MET.

F.2.4. STEP 4: EVIDENCE LINKAGE CHECK

**Criteria Checked:**

- Specific data sources cited: **YES**
  - sensor\_analysis.md (next-day slope estimates; meanRMSSD variability)
  - literature\_review.md (PMID 39903564, 40543159, 41348148)
- Logic from evidence to parameters: **SOUND (with minor documentation gap)**
  - You report that a fitter tool was used to match the stated 90% interval with df fixed at 7.
- Not overly optimistic or conservative: **PASS**
  - Interval crosses 0 and allows both benefit/harm directions.

**My Reasoning:** Evidence linkage is acceptable at the validation level, but the protocol should explicitly define the MET predictor (daily mean MET vs MET-hours; raw vs standardized) because that directly changes the effect-size interpretation.

F.2.5. STEP 5: COMPLETENESS CHECK

- All required fields present: **YES** (distribution, parameters, interval, context, evidence pointers)
- Reasoning sufficiently detailed: **YES**
- Limitations acknowledged: **PARTIAL** (recommend adding explicit note about  $\sigma$  vs SD and predictor scaling)

**F.3. Checklist Summary**

Table 5. Validation Criteria Summary

Criterion	Status	Evidence	Notes
Distribution Type	PASS	Slope is unbounded; Student-t appropriate	Robust heavy tails
Parameter Bounds	PASS	$df > 0$ , $scale > 0$	OK
Clinical Plausibility	PASS	Implied net change across $\Delta MET \approx 1.7$ is moderate	Center slightly negative but interval crosses 0
Evidence Support	PASS	sensor_analysis.md + literature PMIDs	Clarify predictor scaling
Completeness	PASS (minor edits)	Interval + parameters provided	Add $\sigma$ vs SD note

**F.4. Overall Assessment**

**Valid: YES**

**Detailed Reasoning:**

- Internal consistency with the claimed 90% interval is accepted based on your fitter tool output (df=7, loc=-0.5, scale=1.8474; interval match YES).
- Parameterization is appropriate and needs a brief clarifying statement that  $\sigma$  is the Student-t scale, not marginal SD.
- The implied effect sizes over the observed MET range are clinically plausible for meanRMSSD measured in ms.

**F.5. Critique**

Well-constructed robust prior for a regression slope with a clearly stated intended 90% interval. The primary improvement needed is documentation clarity: define the predictor scaling precisely and explicitly distinguish Student-t scale from marginal SD.

**F.6. Suggestions**

1. In the prior table, explicitly state: Student-t(df=7, loc=-0.5, scale=1.8474); **scale is not SD**; marginal SD 2.19 ms/MET.
2. Define the MET predictor precisely (daily mean MET vs MET-hours; raw vs z-scored) and report its typical range/SD in your dataset.
3. Optionally add a one-line clinical plausibility statement: across  $\Delta MET 1.7$ , 90% prior net change [-6.8,+5.1] ms.