Submitted to the Annals of Applied Statistics

VARIANCE AS A PREDICTOR OF HEALTH OUTCOMES: SUBJECT-LEVEL TRAJECTORIES AND VARIABILITY OF SEX HORMONES TO PREDICT BODY FAT CHANGES IN PERI- AND POST-MENOPAUSAL WOMEN

BY IRENA CHEN[1,a], ZHENKE WU[1], SIOBÁN D. HARLOW[2], CARRIE A. KARVONEN-GUTIERREZ[2], MICHELLE M. HOOD[2] AND MICHAEL R. ELLIOTT[1]

[1]*Department of Biostatistics, University of Michigan,* [a]*irena@umich.edu*

[2]*Department of Epidemiology, University of Michigan*

Longitudinal biomarker data and cross-sectional outcomes are routinely collected in modern epidemiology studies, often with the goal of informing tailored early intervention decisions. For example, hormones such as estradiol (E2) and follicle-stimulating hormone (FSH) may predict changes in womens' health during the midlife. Most existing methods focus on constructing predictors from mean marker trajectories. However, subject-level biomarker variability may also provide critical information about disease risks and health outcomes. Current literature does not provide statistical models to investigate such relationships with valid uncertainty quantification. In this paper, we develop a fully Bayesian joint model that estimates subject-level means, variances, and co-variances of multiple longitudinal biomarkers and uses these as predictors to evaluate their respective associations with a cross-sectional health outcome. Simulations demonstrate excellent recovery of true model parameters. The proposed method provides less biased and more efficient estimates, relative to alternative approaches that either ignore subject-level differences in variances or perform two-stage estimation where estimated marker variances are treated as observed. Empowered by the model, analyses of women's health data reveal, for the first time, that larger variability of E2 was associated with slower increases in waist circumference across the menopausal transition.

1. Introduction.

1.1. *Scientific background and motivation.* The menopausal transition is a critical lifestage that can shape women's midlife and long-term health. The US census bureau estimates that by 2050, approximately 47 million women in the U.S. will be aged 45 to 64 years (U.S. Census Bureau). Worldwide, there were 657 million women aged 45–59 in 2021 (Rees et al., 2021), and women are projected to spend more than one-third of their life post-menopause (Mohammadalizadeh Charandabi et al., 2015). Therefore, understanding how the midlife can affect health outcomes is vital for supporting a healthy aging population.

Reproductive aging and the menopausal transition are characterized by well-established patterns of falling levels of estradiol (E2) and rising levels of follicle-stimulation hormone (FSH) (Randolph et al., 2004). In addition to regulating reproductive functionality, E2 and FSH have also been found to be highly associated with risk of adverse health outcomes (Karvonen-Gutierrez and Harlow, 2017; Zaidi et al., 2018). Since E2 regulates adipose tissue, women tend to gain fat mass post-menopause (Colleluori et al., 2018). Fat mass distribution also changes, with body fat shifting to the intraabdominal region during the menopausal transition (Carr, 2003). Excess abdominal fat is one of the symptoms of metabolic syndrome,

Keywords and phrases: Estradiol, Follicle-stimulating hormone, Hamiltonian Monte Carlo, Joint models, Menopause, Subject-level variability, Study of Women's Health Across the Nation (SWAN), Variance component priors.

which can place individuals at higher risk of health conditions such as heart disease, diabetes, and stroke. Waist circumference is a commonly used measure of abdominal fat and previous research suggests that waist circumference may be an important indicator of health risks (Ross et al., 2020; Darsini et al., 2020).

Higher increases in FSH levels are also associated with higher fat mass increases in women undergoing menopause (Sowers et al., 2007). Additionally, FSH appears to be an important predictor of increased adiposity, reduced energy expenditure (Sponton and Kajimura, 2017; Kohrt and Wierman, 2017) and lower lean mass during the postmenopause (Gourlay et al., 2012). This motivates further investigation into how E2 and FSH can jointly predict body mass composition in women. Identifying these associations is important since excess body weight can increase the risk of adverse health outcomes and mortality in midlife women (Stevens et al., 2002)

Growing evidence also suggests that the variability of these hormones may be critical for predicting adverse health outcomes. Gordon et al. (2016) found that higher E2 variability in women over a period of 14 months was predictive of greater depressive symptoms at month 14. Lower levels of FSH variability in perimenopausal and postmenopausal women was strongly associated with reduced risk of hot flash, while changes in individual mean FSH trajectories were not similarly predictive of hot flash risk (Jiang et al., 2015). By understanding how the variability of these biomarkers relate to changes in body mass, we can improve health diagnostics for women and support individualized treatment plans. Despite this emerging work, the majority of current research has still focused on using mean hormone measurements or group based trajectories to predict health outcomes. The existing literature does not account for how individual variabilities or co-variability of E2 and FSH may be related to changes in body mass and waist circumference across the menopausal transition. Thus, there is a dearth of statistical models that properly extract and use these individual hormone trends and variabilities when predicting an outcome.

1.2. *SWAN Dataset.* Our motivating dataset comes from the Study of Women's Health Across the Nation (SWAN), a multi-site US-based longitudinal cohort study that followed women over the menopausal transition (Sowers et al., 2000). The SWAN dataset has made it possible to establish longitudinal associations between hormone trajectories and health outcomes (Park et al., 2017; Sowers et al., 2007), rather than relying on baseline hormone measurements to predict health risks. The comprehensive and longitudinal aspect of this dataset makes it ideal for understanding how individual hormone trends can predict changes in body mass composition.

To be eligible for the SWAN cohort, women had to be between 42-52 years old, had to have had at least one menstrual period and not used reproductive hormones (e.g. hormonal contraceptives or other exogenous hormones) in the past three months prior to enrollment in the study, had to reside in the geographic area of the clinical site, and had to self-identify as White, Black, Chinese, Japanese or Hispanic. Serum E2 and FSH biomarker measurements were collected at baseline and during 13 of the 15 approximately annual follow-up visits, along with other health measurements. Figure 1 shows E2 and FSH measurements collected from SWAN participants, along with a loess curve to estimate the overall population trend. Body composition was measured via dual-energy X-ray absorptiometry (DXA) at five of the seven clinical site visits. Women also completed questionnaires regarding lifestyle and socio-demographic characteristics.

For the fat mass dataset, we initially started with women who were enrolled at one of the five sites with body composition measures. If a woman was on hormone replacement therapy during a clinical visit, we removed that observation from the dataset. Additionally, we removed women who did not have an observed FMP. Although the SWAN study enrolled

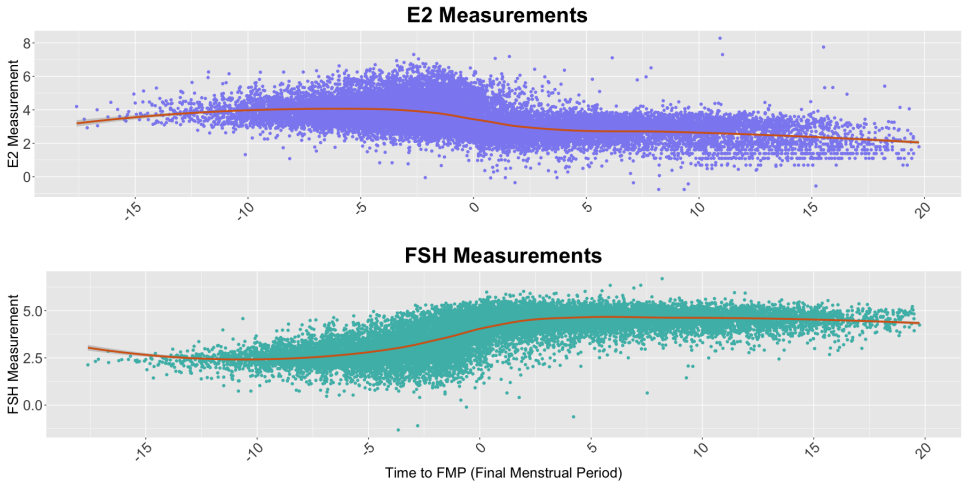

FIG 1. *Plots of the observed E2 and FSH measurements. In our analysis, we log-transformed these measurements and then detrended them by subtracting the individual observations from a population loess fit. See S7 of the Supplementary Materials for additional commentary regarding the loess fit for the FSH measurements.*

five racial/ethnic groups, the site with Hispanic women did not have body composition data; hence, Hispanic women are not included in the fat mass analysis. After computing the fat mass composition window (see below), an additional 47 women were excluded from the analysis, due to either not having both pre and post FMP observations or not having observations that fell within the desired time range before and after FMP (i.e. observations outside of the 3-7 year range before and after FMP). The 3-7 year range was chosen in order to ensure that the changes in fat mass distribution was captured sufficiently before the start (and end) of the menopausal transition, and the number of women who had measurements beyond the 7 year cutpoint before and after FMP was scarce. The final analytical fat mass dataset was completed on 841 individuals, with a total number of 9,902 hormone measurements.

For the individual trajectory model, we use the log values of FSH (mIU/mL) and E2 (pg/mL) measured at each visit, as shown in Figure 1. We removed the E2 and FSH population trends by fitting a lowess curve to each (log) hormone. The lowess curve was fit by using time to FMP at each visit as the numeric predictor for the corresponding hormone measurement and using weighted least squares to obtain the predicted fit at each timepoint. We then subtracted the individual measurements from the lowess estimates. By removing the common population trends in the data, our model can better approximate the individual trajectories and individual level variances using a simpler (lower dimensional) subject-level trajectory model. Figure 3 summarizes the subject-level longitudinal data model fit results for two randomly-selected women.

For waist circumference analysis, since all seven sites collected waist circumference measurements, we were able to have a larger sample size for this analysis. As in the fat mass analysis, we removed women who did not have an observed FMP and we also removed observations where the woman was on hormone replacement therapy. Our final analytical dataset for waist circumference was completed on 1,029 individuals and 12,059 hormone measurements.

For the outcomes of interest, we selected fat mass rate of change and waist circumference rate of change over a selected time window. We define this window to be from the visit closest to 5 years before the FMP to the visit closest to 5 years after the FMP, with the requirement that the closest visit be at least 3 or more years before/after the FMP. By doing this, we aimed to capture the most accurate trend in body composition change that was not fully dependent

Variable	Statistic	Value	n
Longitudinal Predictors	Mean/SD		
E2 Residuals		-0.04 (0.81)	9,902
FSH Residuals		0.02 (0.61)	9,902
Health Outcome	Mean/SD		
Fat Mass Rate of Change		0.001 (0.004)	841
Baseline Body Mass	Mean/SD		
Fat Mass Prop. at Visit 1		0.36 (0.07)	841
Race/Ethnicity	Percent		
White (Reference)		47.2%	397
Black		24.9%	209
Japanese		15.3%	129
Chinese		12.6%	106
Physical Activity	Percent		
Lowest Activity (Reference)		23.6%	199
Increasing Activity		12.7%	107
Decreasing Activity		22.7%	191
Middle Activity		25.6%	215
Highest Activity		15.3%	129

TABLE 1

Descriptive statistics of the fat mass dataset based on 841 individuals.

Variable	Statistic	Value	n
Longitudinal Predictors	Mean/SD		
E2 Residuals		-0.003 (0.80)	12,059
FSH Residuals		-0.008 (0.62)	12,059
Health Outcome	Mean/SD		
Waist Circumference Rate of Change		0.41 (0.82)	1,029
Baseline Waist Circumference	Mean/SD		
Waist Circumference at Visit 1		86.03 (15.90)	1,029
Race/Ethnicity	Percent		
White (Reference)		47.0%	484
Black		26.2 %	270
Japanese		15.3%	128
Chinese		11.5%	119
Hispanic		2.7%	28
Physical Activity	Percent		
Lowest Activity (Reference)		23.6%	267
Increasing Activity		12.7%	131
Decreasing Activity		22.6%	233
Middle Activity		24.0%	247
Highest Activity		14.6%	151

TABLE 2

Descriptive statistics of waist circumference dataset based on 1,029 individuals.

on a measurement right before menopause. The fat mass (waist circumference) rate of change is the difference between the 'last visit' (post FMP) and the 'first visit' (pre FMP) divided by the amount of time (in years) within each individual window. Figure 2 shows the individual observations of these rates. We normalized the fat mass measurements by using the proportion (i.e. ratio) of fat mass to body weight (grams) rather than using the unadjusted fat mass measurements (also in grams), thus creating a measure of percent fat mass for each woman. Raw fat mass values are highly correlated with body weight, and previous work has demonstrated that there is no menopausal effect of body weight change beyond normal aging. The proportion of fat mass, however, has a strong curvilinear relationship across the menopausal transition. For this reason, we have used fat mass proportion in our analyses to reflect the impact of the MT on this measure. (Greendale et al., 2019). Figure 2 displays the histograms of these outcomes, after performing the normalization (for fat mass) and rate adjustments (for both models).

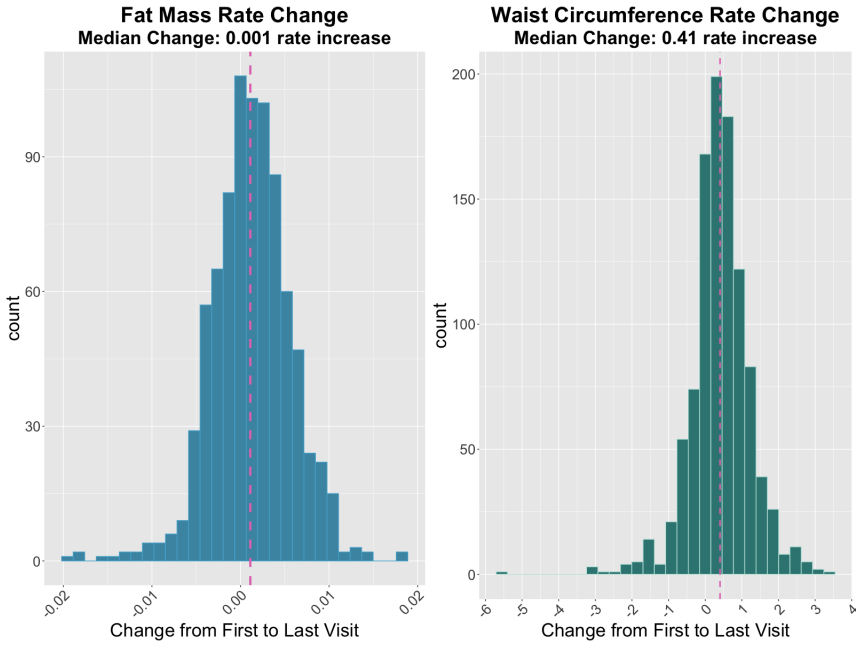

FIG 2. *Histograms of the observed rate of change in fat mass composition (left) and the observed rate of change in waist circumference (right). The observed values for the fat mass outcome are in fat mass proportional to body weight (both in grams) per year and waist circumference (cm) per year.*

1.3. *Methods for longitudinal markers and health outcomes.* A large and well-developed body of literature over the past two decades uses longitudinal biomarker or questionnaire data to predict health outcomes: an early example is given by Henderson, Diggle and Dobson (2000), who tied psychiatric disorder measures over time to predict dropout in schizophrenia trials; they used a subject-level random effect in the disorder trajectory that is also present as a frailty in the time-to-event model. Proust-Lima et al. (2014) link latent classes of prostate specific antigen to survival models. More recent work by Wang, Luo and Li (2017) considers multiple longitudinal predictors - in their case, measures of daily functioning - to predict onset of Parkinson's disorder.

However, methods that assess the utility of residual variability in predicting health outcomes are largely lacking in the literature, despite calls for increased focus on developing methods for variance structures. Such methods may elucidate whether variability as predictors could yield both better prediction and improved inference (Carroll, 2003). This oversight appears to be a substantial gap in the statistical methods literature given that increased variance can be an early predictor of instability in biological systems; for example, heart rate variability may be a marker of autonomic dysregulation given its predictive nature with poor health outcomes (Young and Benton, 2018). Our proposed method addresses this gap by explicitly parameterizing individual variances (and co-variances) and uses these estimates as predictors in the outcome model.

Early literature used simple two-stage approaches with a squared-error estimate of variance obtained from the observed data to predict a single outcome (Sammel et al., 2001), ignoring the inherent uncertainty in the constructed variance estimates. More recent methods have focused on a joint model for the predictors and outcomes (Elliott, Sammel and Faul, 2012; Jiang et al., 2015, e.g.,). The joint modeling approach is critical given that statistical uncertainty of the constructed predictors, e.g., mean and residual variance estimates, if unaccounted for, can lead to extremely biased estimates of the effects of these individual-level mean trajectories and the residual variances on the outcomes (e.g., Ogburn et al., 2021;

Wang, McCormick and Leek, 2020). Finally, prediction approaches based on multiple trajectories - which allow for consideration of residual covariances as well as residual variances as predictors – appears to be completely absent from the biostatistics toolkit. In a Bayesian framework, our joint model properly accounts for the uncertainty in estimating the residual variances and covariances of multiple biomarkers, thereby propagating the uncertainty into outcome prediction.

This paper is organized as follows. Section 3 describes our joint model framework linking the mean trajectory and residual variance model for the longitudinal predictors with the model for the outcome. Section 5 conducts extensive simulation studies to show that our proposed approach produces less biased estimates of the outcome regression coefficients with valid statistical uncertainty assumptions relative to a variety of increasingly complex two-stage competitors that fail to account for the statistical uncertainty in the subject-level means and in the variances. Section 6 applies our method to assess the associations between the mean trajectories, variance, and covariance of E2 and FSH and the changes in fat body mass and fat distribution during the menopausal transition using women's health data from the SWAN study. Section 7 discusses the implications of our work along with limitations and directions for future research.

2. Previous Work on Variances as a Predictor of Outcomes. Joint modeling of longitudinal trajectories and cross-sectional outcomes is a rich area of statistical research (Chi and Ibrahim, 2006; Ibrahim, Chu and Chen, 2010; Lawrence Gould et al., 2015). For example, including longitudinal measurements in joint longitudinal - survival models can provide better estimates of the survival outcome by accounting for the individual trajectories over time, as shown by Long and Mills (2018), who used a joint model to predict motor diagnosis using longitudinal characterizations of Huntington's disease progression taken at annual visits; also see Papageorgiou et al. (2019) for a comprehensive review. Until recently, most of the focus in the area of modeling individual trajectories has been on modeling the mean trends. The variability associated with individual biomarkers has traditionally been treated as a nuisance parameter. In mathematical notation, these joint models typically take the form:

$$\boldsymbol{X}_{ij} \mid \boldsymbol{B}_i, \Psi \sim \mathcal{N}(\mu(\boldsymbol{B}_i, t_{ij}), \Psi),$$
$$Y_i \mid \boldsymbol{B}_i, \beta, \sigma^2 \sim \mathcal{N}(\eta(\beta, \boldsymbol{B}_i), \sigma^2),$$

where X are the observed markers/predictors, Y is the observed outcome of interest, the main predictors of interest are the mean parameters $\boldsymbol{B}_i$ and the variance-covariance matrix Ψ is assumed to be a population level parameter, rather than allowing each individual to have their own variance-covariance matrix (e.g. S_i).

However, interest in both modeling the residual variability of individual trajectories and using these estimates as predictive variables has been growing. Elliott, Sammel and Faul (2012) studied the relationship between individual variability in short-term memory tests and long term onset of senility. They found that increased variability in the memory tests were associated with increased risk of senility. Furthermore, this variability was a stronger predictor of senility than the mean trajectories. With regards to women's health, some previous research has considered variance as a predictor of health outcomes. Harlow, Lin and Ho (2000) found that women who had increased menstrual cycle variability at a younger age were more likely to experience abnormal uterine bleeding. The variability of menstrual cycle length was also found to be an important predictor of menopausal onset (Huang, Elliott and Harlow, 2014). Sammel et al. (2001) used a two-stage model to that linked individual means and variances of longitudinal profiles to a corresponding health outcome. They found that E2 variability during the menopausal transition was highly predictive of experiencing

hot flashes. Jiang et al. (2015) proposed a joint model of individual means and variances of FSH hormone trajectories and risk of hot flash as the health outcome. Low FSH variability was predictive of substantial reduction in risk of hot flash. We note that a common feature of these models is the lack of multiple predictors in the longitudinal sub-model.

Our work's most important contribution over the previous work is to consider the individual-level means and variances of multiple predictor trajectories, rather than a single biomarker trajectory, in a joint modeling framework. This allows for investigation into how these trajectories may both independently and interactively associate with outcomes, and in turn requires substantial methodological development, particularly to decompose an individual-level variance-covariance matrix for use in a joint modeling setting. This work also adds to the still small set of literature showing the use of variability as a predictor of health outcomes, and more generally emphasizes the need for a joint modeling framework over the less efficient two-stage approach, as we show in Section 5. Also, to the best of our knowledge this is the first rigorous assessment of the role of individual trends and variability of E2 and FSH hormones in jointly predicting body mass changes.

2.1. *Measurement Error in Multivariate Linear Models.* An advantage of joint models relative to two-stage models is that the uncertainty associated with the parameters estimated in the first stage is carried over to the second stage. Consider a simple linear relationship of the form:

$$Y = X\beta + \epsilon, \tag{1}$$

where X is a $n \times K$ matrix of K predictors and ϵ is an $n \times 1$ vector of independent normal error terms with mean 0 and variance σ^2. Suppose that the true relationship between Y and X is described by (1), but instead, we observe $\widetilde{X}$, where $\widetilde{X} = X + U$, where U is the matrix of normally-distributed independent measurement errors with mean 0 and variance-covariance Σ_U. If $U \perp\!\!\!\perp X$, then in the $K = 1$ scenario, we know that the estimate of β will be attenuated towards the null (Carroll et al., 2006, p.42-43).

In the Bayesian setting, given $\widetilde{X} = X + U$ where $\widetilde{X}$ is the observed predictor, X is the (unknown) predictor, and U is the measurement error, the bias is now also dependent on the choice of prior for $\widetilde{X}$. With a sufficiently weak prior choice (a weak Gaussian prior is suggested in Stan Development Team (2023) and enough data, the bias from a two-stage Bayesian model should be similar to the bias from a frequentist two-stage (Richardson and Gilks; Bartlett and Keogh). In the limit as n grows large, we expect the biases of the frequentist and Bayesian approaches to converge.

For $K > 1$, with multiple predictors measured with error, the estimates of the β are still biased, but the direction of the bias now depends on the correlation between the measurement errors (Carroll et al., 2006, p.53-55). Consider the following equation for $K = 2$ predictors:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{2}$$

and suppose we measure X_1, X_2 with some error:

$$\widetilde{X}_1 = X_1 + U_1, \widetilde{X}_1 = X_2 + U_2. \tag{3}$$

Griliches and Intriligator (1987, p.1477–1479) derive the bias of estimating β_1 as:

$$\textit{plim}(\widehat{b}_1 - \beta_1) = -\frac{\beta_1 \lambda_1}{(1-\rho^2)} + \frac{\beta_2 \lambda_2 \rho}{(1-\rho^2)}, \tag{4}$$

where $\widehat{b}_1$ is the coefficient obtained from regressing Y on $\widetilde{X}_1$ in the multiple regression model, $\lambda_1 = \mathrm{Var}(U_1)/\sigma_1^2$ is the relative amount of measurement error in X_1 where σ_1^2 refers

to the variance of X_1, and ρ is the (true) correlation between X_1, X_2; *plim* refers to convergence in probability. A similar equation can be derived for the bias of estimating β_2 in the presence of such measurement errors. We can see, then, that the bias is increased by a factor of $\frac{1}{(1-\rho^2)}$. The overall effect of the additional variable $\widetilde{X}_2$ is a bias towards the null (Griliches and Intriligator, 1987, p.1479).

For $K > 2$ variables, the expressions for the bias of each predictor become more complicated to derive. We show via simulations in Section 4 that the bias in the mean parameters is clear in the two-stage model linear regression alternatives to the joint model. Furthermore, we see that that bias in the estimates of the variance-covariance parameters persists in the two-stage model alternatives. This would be an issue if individual variability (covariability) is predictive of an outcome of interest.

3. Proposed Model.

3.1. *Notation.* Let $\mathcal{D} = (Y_i, \boldsymbol{X}_{ij}, t_{ij}, \boldsymbol{W}_i)$ be the observed data for subject $i = 1, \ldots, N$, where $Y_i \in \mathbb{R}^1$ is a continuous outcome, $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijQ})^\mathsf{T}$ is a vector of Q time-varying marker values at observation time points $t_{ij}, j = 1, \ldots, n_i$, that may differ by subjects, and $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{id})^\mathsf{T}$ is a vector of d time-invariant covariates, e.g., race/ethnicity, activity class.

The proposed model has two connected components. We first specify the regression model for irregularly and longitudinally observed multiple markers $(t_{ij}, \boldsymbol{X}_{ij}), j = 1, \ldots, n_{ij}$; the second component links the outcome Y_i to time invariant covariates $\boldsymbol{W}_i$ and unobserved individual-specific vectors of regression coefficients and residual variance-covariance matrices in the model for longitudinal markers, enabling inference about how the mean trajectories and residual variations are associated with the outcome Y_i.

3.2. *Likelihood.*

3.2.1. *Component 1: Longitudinal Markers.* We specify the model for the longitudinal marker data as follows:

$$\boldsymbol{X}_{ij} \mid \mathbf{B}_i, \mathbf{S}_i \sim \mathcal{N}_Q\left(\boldsymbol{\mu}(t_{ij}; \mathbf{B}_i), \mathbf{S}_i\right), j = 1, \ldots, n_i, \text{ independently for } i = 1, \ldots, N, \tag{5}$$

$$\boldsymbol{b}_{iq} \overset{\text{indep.}}{\sim} \mathcal{N}_P(\boldsymbol{\beta}_q, \Sigma_q), q = 1, \ldots, Q, \tag{6}$$

where $\mathcal{N}_Q(\boldsymbol{\mu}, \mathbf{S})$ is a generic notation that represents a Q-dimensional multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{S}$; $\boldsymbol{\mu}(t; \mathbf{B}_i)$ is a Q-dimensional function of time given by $\mathbf{B}_i = [\boldsymbol{b}_{i1}, ..., \boldsymbol{b}_{iQ}]^\mathsf{T}$ and $\boldsymbol{b}_{iq} = (b_{iq1}, \ldots, b_{iqP})^\mathsf{T}$ is a vector of P regression coefficients for the q-th marker. Here P is the number of basis functions of time; to simplify presentation in this paper, we assume the same number of coefficients for each marker, e.g., intercept and slope. In our simulation and data application, we specify these basis functions in advance to be linear functions of time. In addition, $\boldsymbol{\beta}_q = (\beta_{q1}, \ldots, \beta_{qP})^\mathsf{T}$ is a vector of population mean regression coefficients that are specific to the q-th marker. In addition, Equation (6) has assumed individual-specific regression coefficients for any two markers are conditionally independent. Note that in this model, we have given each individual a variance covariance matrix $\mathbf{S}_i$.

3.2.2. *Component 2: Outcome Regression (OR) Model.* The outcome variable Y_i is assumed to be related to individual-specific mean and variance-covariance parameters $\mathbf{B}_i$ and $\mathbf{S}_i$ in the longitudinal marker model (5) as follows:

$$Y_i \mid \mathbf{B}_i, \mathbf{S}_i, \boldsymbol{W}_i \sim \mathcal{N}\left(\eta_i, \sigma^2\right), i = 1, \ldots, N, \tag{7}$$

where $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{id})^\mathsf{T}$ is a vector of d time-invariant covariates; $\eta_i = \eta(\mathbf{B}_i, \mathbf{S}_i, \boldsymbol{W}_i; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\gamma}^W)$ and $\eta(\cdot; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\gamma}^W)$ is a generic mean outcome regression parameterized by $\boldsymbol{\alpha}$ (for $\mathbf{B}_i$), $\boldsymbol{\gamma}$ (for $\mathbf{S}_i$), and $\boldsymbol{\gamma}^W$ (for $\boldsymbol{W}_i$); $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\gamma}^W$ are of dimension PQ, $Q(Q+1)/2$, d, respectively. In this paper, we will illustrate the statistical performance of such a formulation by focusing on simple specifications of $\eta(\cdot)$, e.g., linear models. The framework readily generalizes to general outcome models; in Section 6.2, we illustrate a model with a Gaussian scale-mixture outcome regression model for waist circumference rate-of-change outcome.

3.3. *Priors.* In this section, we specify the prior and hyperprior distributions for the unknown parameters in the two likelihood components.

Prior for $\mathbf{S}_i$. We rewrite $\mathbf{S}_i = \mathbf{D}_i \mathbf{R}_i \mathbf{D}_i$, where $\mathbf{D}_i = \text{diag}(d_{i1}, ..., d_{iQ})$ is a diagonal matrix of residual variances and $\mathbf{R}_i$ is the associated subject-level correlation matrix. We assume

$$\log(d_{iq}) \sim \mathcal{N}(\nu_q, \psi_q^2), \tag{8}$$

independently for marker $q = 1, \ldots, Q$, and subject $i = 1, \ldots, N$. Because $\mathbf{R}_i$ is a correlation matrix, we only need to specify the prior distribution for the off-diagonal elements. We consider the special case of $Q = 2$, where r_{12i} is unconstrained, and separately the general case of $Q > 2$, where the components of $\mathbf{R}_i$ must meet the positive definite criterion.

Special Case: $Q = 2$.. For a 2×2 correlation matrix, we place the following prior on the off-diagonal value r_{12}:

$$(r_{i12} + 1)/2 \sim \mathsf{Beta}(a'_{12}, b'_{12}), \text{ independently for } i = 1, \ldots, N. \tag{9}$$

Finally, we specify hyperpriors for ν_q, ψ_q and (a'_{12}, b'_{12}) by

$$\nu_q \sim \mathcal{N}(m, \xi_q^2), \psi_q \sim \mathsf{half\text{-}Cauchy}(0, \tau), \text{ independently for } q \leq Q, \tag{10}$$

$$a'_{12} \sim \mathsf{Exp}(\kappa),\ b'_{12} \sim \mathsf{Exp}(\kappa'). \tag{11}$$

In our application, we set $m = 0$ and $\xi_q = 10$. In addition, we use a half-Cauchy hyperprior on ψ_q instead of the inverse-Gamma distribution as this prior is recommended for datasets where the signal of the variance ψ_q may be weak (Gelman, 2006). In this setting, inferences using the inverse-Gamma distribution are extremely sensitive to the choice of hyperparameter values (Gelman, 2006, p. 524), which makes the inverse-Gamma prior "not at all uninformative". The half-Cauchy distribution avoids this potential issue due to its heavier tail, which still allows for higher estimates of the variance, but constrains the posterior distribution "to an extent allowed by the data".

REMARK 1. *The defining feature of our framework is the individual-specific variance-covariance matrices,* $\{\boldsymbol{S}_i, i = 1, \ldots, N\}$*, over which we must specify a hierarchical prior distribution. Such priors must not be restrictive in capturing between-subject differences and similarities in the variance-covariance matrices. Focusing on the prior specification for individual-specific correlation matrices* $\{\mathbf{R}_i, i = 1, \ldots, N\}$*, standard priors designed for a single unknown population correlation matrix, e.g., Lewandowski-Kurowicka-Joe (LKJ) prior (Lewandowski, Kurowicka and Joe, 2009), have severe drawbacks. In particular, the LKJ distribution is governed by a single positive scale parameter,* ζ*, that tunes the strengths of the correlations. The off-diagonal elements of a* $K \times K$ *correlation matrix are marginally distributed as:* $(r_{ilk} + 1)/2 \sim \mathsf{Beta}(\mathsf{a}, \mathsf{b})$*, where* $a = b = \zeta - 1 + K/2$*. This implies that the correlations will a priori be concentrated around* 0*. However, in our motivating application,* $\{r_{ilk}, i = 1, \ldots, N\}$ *represent the individual-specific residual correlations between the* l*- and*

the k-th hormone, which 1) by domain knowledge are a priori unlikely to have a strong prior of being near zero, and 2) may vary between subjects in a way far from the implied distribution of $2 \cdot \mathsf{Beta}(a, b) - 1$. In Equation (11), we have removed this identity restriction and specified hyperpriors on a and b, which provides greater flexibility in allowing the data to estimate the true a and b. The same argument can be applied to using Inverse-Wishart distribution as a prior for variance-covariance matrices, which is also governed by a single scale parameter and suffers from the same drawbacks as the LKJ distribution.

There is little existing literature on hyperprior recommendations for the parameters of the Beta distribution. Robert and Casella (2010) note that "there exists a family of conjugate priors on a, b", however, they also note that these prior distributions are often intractable, due to the "difficulty of dealing with the Gamma function". Instead, we opt for a simpler approach by allowing the a, b parameters to be independently drawn from an Exponential prior. We argue that the Exponential distribution in Equation (11) is a reasonable choice for a hyperprior as follows: Let $x_1, \ldots, x_n$ be data from a $\mathsf{Beta}(a, b)$ distribution. Assume that $a \sim \mathsf{Exp}(\lambda_a), b \sim \mathsf{Exp}(\lambda_b)$. We can also assume a and b are independent *a priori*. Then the posterior distribution, $p(a, b|x) \propto L(a, b)p(a)p(b) \propto \prod_{i=1}^{n} \exp(-\ln B(a, b) + (a-1)\ln x_i + (b-1)\ln(1-x_i)) \times \exp(\ln \lambda_a - \lambda_a a + \ln \lambda_b - \lambda_b b) \propto \exp(-n \ln B(a, b) + a[\sum_{i=1}^{n} \ln x_i - \lambda_a] + b[\sum_{i=1}^{n} \ln(1-x_i) - \lambda_b])$, which suggests that the posterior distribution of a, b would be updated from a flat prior by subtracting (λ_a, λ_b) from the sufficient statistics $\sum_{i=1}^{n} \ln x_i$ and $\sum_{i=1}^{n} \ln(1-x_i)$. The reason for this is that the Maximum Likelihood Estimators (MLEs) for a, b respectively can be approximated by: $\widehat{a}_{MLE} = \frac{1}{2} + \frac{\widehat{G}_x}{1 - \widehat{G}_x - \widehat{G}_{(1-x)}}, \widehat{b}_{MLE} = \frac{1}{2} + \frac{\widehat{G}_{(1-x)}}{1 - \widehat{G}_x - \widehat{G}_{(1-x)}}$ where $\widehat{G}_x = \exp(n^{-1} \sum_{i=1}^{n} \ln(x_i)), \widehat{G}_{1-x} = \exp(n^{-1} \sum_{i=1}^{n} \ln(1 - x_i))$, where $(a, b) > 1$. The posterior modes using the Exponential priors become $\frac{1}{2} + \frac{\widehat{G}_x \exp(\lambda_a)}{1 - \widehat{G}_x \exp(\lambda_a) - \widehat{G}_{(1-x)} \exp(\lambda_a)}, \frac{1}{2} + \frac{\widehat{G}_{(1-x)} \exp(\lambda_b)}{1 - \widehat{G}_x \exp(\lambda_b) - \widehat{G}_{(1-x)} \exp(\lambda_b)}$. When $\lambda_a, \lambda_b \to \infty$, the posterior modes of a, b shrink towards $\frac{1}{2}$, which is the Jeffrey's prior. When $\lambda_a, \lambda_b \to 0$, we recover the likelihood. Overall, these results suggest that the choice of the Exponential distribution is a flexible hyperprior on a, b and thus is a reasonable choice.

General Case: $Q \geq 3$.. For the general case of a $Q \times Q$ correlation matrix where $Q \geq 3$, the off-diagonal values of the individual correlation matrices are now more complicated to estimate since the space of valid correlation matrices is a proper subset of the space of all possible $Q \times Q$ matrices. We address this constraint by following the approach of Ghosh, Mallick and Pourahmadi (2021) where the off-diagonal values are parameterized in terms of hyperspherical coordinates. The angles are allowed to vary freely over $[0, \pi]$ before being back-transformed into valid correlation values. To illustrate, we specify the prior for $\mathbf{R}_i$ for when $Q = 3$ as follows (similarly for $Q > 3$):

$$r_{12} = \cos(\theta_{12}), r_{13} = \cos(\theta_{13}), r_{23} = \sin(\theta_{12}) \cdot \sin(\theta_{13}) \cdot \cos(\theta_{23}) + \cos(\theta_{12}) \cdot \cos(\theta_{13}),$$

where $\theta_{12} = \arccos(c_{12}), \theta_{13} = \arccos(c_{13}), \theta_{23} = \arccos(c_{23})$, and

$$(c_{i12} + 1)/2 \sim \mathsf{Beta}(a'_{12}, b'_{12}), (c_{i13} + 1)/2 \sim \mathsf{Beta}(a'_{13}, b'_{13}), (c_{i23} + 1)/2 \sim \mathsf{Beta}(a'_{23}, b'_{23}).$$

As in the case where $Q = 2$, we specify hyper-priors for $a'_{kl}, b'_{kl}, k < l$, e.g., the Exponential prior.

3.3.1. *Priors for population longitudinal marker regression coefficients:.*

$$\boldsymbol{\beta}_q \sim \mathcal{N}_P(0, \xi_q^2 I_{P\times P}), \text{ independently for } q = 1, \ldots, Q \tag{12}$$

$$\Sigma_q = \mathbf{K}_q \mathbf{L}_q \mathbf{K}_q, \ \mathbf{K}_q = \text{diag}\{k_{q1}, \ldots, k_{qP}\}, q = 1, \ldots, Q, \tag{13}$$

$$k_{qp} \sim \text{half-Cauchy}(0, \tau_0), p = 1, \ldots, P, \text{ and } \mathbf{L}_q \sim \text{LKJ}(\zeta), \tag{14}$$

independently for $q = 1, \ldots, Q$ where $\mathbf{K}_q = \text{diag}\{k_{q1}, \ldots, k_{qP}\}$ is a diagonal matrix and $\mathbf{L}_q$ is a correlation matrix. The τ_0, ζ parameters are set in practice as 2.5 and 1 respectively. It is fine to use the half-Cauchy and the LKJ priors in Equation (14) since, for each marker q, they are standard hyperpriors for a *single* population variance matrix $\mathbf{K}_q$ and a *single* population-level correlation matrix $\mathbf{L}_q$, which is different from Equations (8 - 9) that specifies the prior for multiple and individual-specific variance-covariance matrices.

3.3.2. *Prior for parameters in the outcome regression model.* For the outcome model, we place diffuse independent Gaussian priors for each element of the outcome regression parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^W$). Finally, to complete the prior specification, we place a diffuse prior on the outcome residual standard deviation parameter $\sigma \sim \text{half-Cauchy}(0, \tau_1)$. In our simulation studies and data analysis, we set the priors on the regression parameters as $\mathcal{N}(0, 10^2)$ (a weakly informative prior in order to allow the data to estimate the parameters) and $\tau_1 = 2.5$ (the default suggested by Carpenter et al. (2017)).

Let $Z = (\boldsymbol{B}_i, \boldsymbol{S}_i)$ and let $\Theta = (\boldsymbol{\beta}_q, \Sigma_q, \xi, \nu_q, \psi_q, a'_{kl}, b'_{kl}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\gamma}^W, \sigma)$, where these sets denote the unknown parameters of interest in the proposed model. Let $\pi(\Theta)$ denote the prior distribution where we have assumed that all parameters in Θ have independent components: $\pi(\Theta) = \Pi_{q=1}^{Q}[\pi(\boldsymbol{\beta}_q)\pi(\Sigma_q)\pi(\xi_q)\pi(\nu_q, \psi_q)]\Pi_{k<l}[\pi(a'_{kl}, b'_{kl})]\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\gamma}^W))\pi(\sigma)$.

3.3.2.1. Joint Distribution. The joint distribution of the data and unknown parameters is then

$$\begin{aligned} P(\Theta, Z, \mathcal{D}) \propto \prod_{i=1}^{n}\prod_{q=1}^{Q}\Big\{ & \frac{1}{\sqrt{(2\pi)^q|\Sigma|}}\exp(-\frac{1}{2}(\boldsymbol{b}_{iq} - \boldsymbol{\beta}_q)^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{b}_{iq} - \boldsymbol{\beta}_q)) \\ & \times \frac{1}{\sqrt{2\pi\xi_q^2}}\exp\left[\frac{(\log(d_{iq}) - \nu_q)^2}{2\xi_q^2}\right]\left(\frac{[(r_{ikl}+1)/2]^{a'_{kl}-1}\{1 - ([(r_{ikl}+1)/2]\}^{b'_{kl}-1}}{\text{Beta}(a'_{kl}, b'_{kl})}\right) \\ & \times \prod_{j=1}^{n_i}\frac{1}{\sqrt{(2\pi)|\mathbf{S}_i|}}\exp\left(-\frac{1}{2}\{\boldsymbol{X}_{ij} - \boldsymbol{\mu}(t_{ij}; \mathbf{B}_i)\}^{\mathsf{T}}\mathbf{S}_i^{-1}\{\boldsymbol{X}_{ij} - \boldsymbol{\mu}(t_{ij}; \mathbf{B}_i)\}\right)\Big\} \\ & \times \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{(Y_i - \eta(\mathbf{B}_i, \mathbf{S}_i, \boldsymbol{W}_i; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\gamma}^W))^2}{2\sigma^2}\right] \times \pi(\Theta). \end{aligned} \tag{15}$$

Figure S1 in the Supplementary Materials uses a directed acyclic graph to visualize and summarize the hierarchical relationships between the different components of our modeling framework.

3.4. *Posterior Inference.* In a Bayesian framework, the inference is conducted based on the posterior distribution $P(\Theta \mid \mathcal{D})$. However, it is not feasible to derive the closed-form posterior distribution owing to the lack of prior-likelihood conjugacy in our proposed model. We therefore used Hamiltonian Monte Carlo to draw sequential samples and approximate the posterior distribution. We implement the model using Stan and the `rstan` package (Stan Development Team, 2020) as the interface for running the model and obtaining the posterior estimates. Code to run the joint model and generate the data used in our simulation studies

are provided in attached supplementary files. For our simulations studies in Section 5.1, 5.2 and Section 3 in the Supplementary Material, we run two chains per independent replicate data set, with $2,000$ iterations and $1,000$ burn-in. For the data application in Section 6, we ran 4 chains each for 2,000 iterations with 1,000 burn-in. Visual inspection of the traceplots for all model parameters indicated non-divergent chains (see Figures S4 and S5 in the Supplementary Materials). All chains were combined for calculating posterior summaries.

We also examined Stan's R-hat convergence diagnostic (Vehtari et al., 2021) and the Effective Sample Sizes (ESS) to determine if the chains had mixed well. The R-hat value for all model parameters was less than 1.05. In the fat mass rate of change model, the ESS for all of the model parameters was at least 100 times the number of chains used, except for the model parameter corresponding to individual-level E2 variance and the parameter corresponding to fat mass proportion at the first visit. The R-hat values of these two parameters was also effectively 1.00 in both models. Based on these diagnostics, we concluded that our models had converged. The posterior predictive checks we conducted (see S2 in the Supplementary Material) suggest that our model generates reasonable estimates for the observed outcomes and trajectories.

4. Alternative Methods. We briefly introduce three common alternatives in our comparative simulation study: two-stage simple linear model (TSLM), two-stage linear mixed model (TSLMM), and two-stage individual-variance (TSIV) model. We refer to our joint model as the "Joint Model with Individual Variances", or JMIV. We compare the performance of JMIV to these alternative methods in Section 5.

4.0.0.1. Two-Stage Simple Linear Regression (TSLM). One of the most simple alternative models we could use is the linear regression model in two stages. We used the `lm()` function in R and first fit the following model:

$$X_{ijq} = \beta_{iq0} + \beta_{iq1} t_{ij} + \epsilon_{iq}, q = 1, 2.$$

Here, we will obtain $\widehat{\beta}_{iq0}, \widehat{\beta}_{iq1}$ via ordinary least squares estimates for the mean parameters b_{iq0}, b_{iq1}. To estimate s_{i11}, s_{i22}, we collected the residuals from each regression, e.g., $r_{ij1} = (x_{ij1} - (\widehat{\beta}_{i10} + \widehat{\beta}_{i11} t_{ij})), j = 1, \ldots, n_i$, and computed the sample variance of these residuals, which we term "estimated residual variance" for each individual, i.e., $\widehat{s}_{i11}$; similarly, we obtain r_{ij2} and $\widehat{s}_{i22}$. The residual covariance, $\widehat{s}_{i12}$, was estimated by sample covariance of (r_{ij1}, r_{ij2}). We then used linear regression to model the outcome based on the estimated coefficients and residual variances and covariances from the first-stage model: $\mathbb{E}(Y_i \mid \text{others}) = \sum_{q=1,2} \left\{ \alpha_{q1} \widehat{\beta}_{iq0} + \alpha_{q2} \widehat{\beta}_{iq1} + \gamma_{qq} \widehat{s}_{qq} \right\} + \sum_{q' < q} \gamma_{q'q} \widehat{s}_{q'q}$. We then performed a bootstrap procedure on the point estimates of each outcome parameter as follows: for $i = 1, \ldots, 1000$:

1. Resample the dataset with replacement
2. Run the outcome linear regression model described above
3. Save the estimated model coefficients

For each replicate, we saved the mean of the 1,000 point estimates as well as the 2.5th and 97.5th quantiles of the 1,000 point estimates. We then used these 1,000 means and 95% "CIs" to compute the bias, coverage, and average interval length.

4.0.0.2. Two-Stage Linear Mixed Model and Linear Regression (TSLMM). This alternative is a slightly more sophisticated approach than TSLM. In the first stage, we fit a Bayesian bivariate response linear mixed model with the **brms** package (Bürkner, 2017)

$$X_{ijq} = \beta_{q0} + b_{iq0} + \beta_{q1} t_{ij} + b_{iq1} t_{ij} + \epsilon_{ijq}, q = 1, 2.$$

We chose to use a Bayesian framework for this model since fitting linear mixed models with multivariate outcomes is more complicated to implement in a frequentist setting. Standard Bayesian software such as the **brms** package allows for easier implementation of multivariate outcome linear mixed models. We place independent $\mathcal{N}(0, 10^2)$ priors on the intercept and slope parameters. We use the preset prior distribution for the random-effects correlation matrix, an LKJ prior with scale parameter 1, as suggested by Bürkner (2017). For all other prior specifications, we used the default prior settings in the **brms** package.

We approximated the "B_i" coefficients for each individual trajectory with the "overall" coefficient estimates: $\widehat{B}_{iq0} = \widehat{\beta}_{q0} + \widehat{b}_{iq0}$ and $\widehat{B}_{iq1} = \widehat{\beta}_{q1} + \widehat{b}_{iq1}$, where $\widehat{\beta}_{qp}$ and $\widehat{b}_{iqp}, p = 0, 1$ are the estimated posterior means of the fixed and random effects respectively. As in the previous model, we estimated $\mathbf{S}_i$ by computing the model residuals (e.g. $X_{ijq} - (\widehat{B}_{i0q} + \widehat{B}_{i1q} t_{ij})$) and then computed the variance across all residuals. We also computed the residual covariance to estimate of $\widehat{s}_{12}$. We then fit the same second-stage model as in the TSLM setup to get the estimated posterior means and corresponding 95% confidence intervals for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

4.0.0.3. Two-Stage Individual Variances (TSIV) Model. Here, we fit the longitudinal outcome model using Equations (5) and (6) only (together with their prior specifications in Equation (9), (11) to (12)) and use the estimates of the posterior means, $\widehat{\mathbf{B}}_i$ and $\widehat{\mathbf{S}}_i$ in the model 7 (together with prior specifications 3.3.2). Note that we do not consider this to be a practical alternative to our first two models, since if one goes to the effort of using a non-standard multilevel model for subject-specific variance-covariance matrices, one might as well go the extra step of bringing them together within a joint model. However, we do this to investigate the effect of not propagating the statistical uncertainty across the two components of the model.

5. Simulation Study. In this section, we present the results from simulation studies with $Q = 2$ and $Q = 3$ biomarkers. Additionally, in Section S3 in the Supplementary Material, we present a simulation study examining our model performance when we approximate the true nonlinear relationship between the marker means and variances and the outcome as linear.

5.1. *Simulation 1: Two Biomarkers.*. In this simulation, we assume the mean trajectories can be expressed linearly with individual intercepts and slopes. We generated $n_i = 6$ to 15 time points for $N = 1,000$ individuals, which mimics the data used in Section 6. We then simulated two trajectories for each individual using the following parameters.

Component 1: Longitudinal Markers .

$$X_{itq} = b_{iq1} + b_{iq2} t + \epsilon_{iq}, q = 1, 2; \boldsymbol{B}_{i1} \sim \mathcal{N}_2\left(\boldsymbol{\beta}_1, \Sigma_1\right), \boldsymbol{B}_{i2} \sim \mathcal{N}_2\left(\boldsymbol{\beta}_2, \Sigma_2\right)$$

$$;\boldsymbol{\beta}_1 = (0, 2)^{\mathsf{T}}, \boldsymbol{\beta}_2 = (2, 1)^{\mathsf{T}}, \Sigma_1 = \begin{pmatrix} 1 & -0.05 \\ -0.05 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & -0.1 \\ -0.1 & 0.5 \end{pmatrix};$$

$$\log(d_{i1}) \sim \mathcal{N}(0, 0.75)/2, \log(d_{i2}) \sim \mathcal{N}(0.5, 0.5)/2, (r_{i12} + 1)/2 \sim \mathsf{Beta}(1, 5).$$

Component 2: Outcome Regression Model. To generate the outcome for each individual, we assume $Y_i \sim \mathcal{N}(\eta(\mathbf{B}_i, \mathbf{S}_i), \sigma^2)$ and set

$$\eta(\mathbf{B}_i, \mathbf{S}_i) = \alpha_{11} b_{i11} + \alpha_{12} b_{i12} + \alpha_{21} b_{i21} + \alpha_{22} b_{i22} + \gamma_{11} s_{i11} + \gamma_{21} s_{i21} + \gamma_{22} s_{i22},$$

where the true values of $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ are $\boldsymbol{\alpha} = (-3, -3, -3, 3), \boldsymbol{\gamma} = (2, -1, 2)$; we did not include other time-invariant covariates $\boldsymbol{W}_i$. These particular truth values were chosen so that the distribution of the outcome Y_i would be similar to the distribution of the SWAN body mass outcomes (our data analysis application). Lastly, we set $\sigma^2 = 1$. We present the results for $R = 200$ replicates in Table 3 for the outcome submodel parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$. See Table S5 in the Supplementary Materials for the results for the other model parameters.

5.1.1. *Simulation I: Results.* Table 3 presents the results of Simulation I. For the two-stage linear regression model, we can see that the point estimates of the outcome model parameters are attenuated towards the null. This result makes sense given what we know about bias resulting from measurement error (Section 2.1). Furthermore, the actual coverage rate is quite poor, especially for the regression coefficients of the variances and covariances (γ).

Truth	Model	Bias	Coverage (%)	Average Interval Length
α_{11} = -3	**JMIV**	0.01	98.0	0.29
	TSLM	0.30	8.0	0.37
	TSLMM	-0.01	93.5	0.38
	TSIV	-0.02	97.5	0.35
α_{12} = -3	**JMIV**	0.01	95.0	0.27
	TSLM	-0.07	83.0	0.31
	TSLMM	0.00	94.0	0.34
	TSIV	-0.01	93.0	0.29
α_{21} = -3	**JMIV**	0.00	97.0	0.25
	TSLM	0.46	0.0	0.29
	TSLMM	-0.01	92.5	0.40
	TSIV	-0.01	89.0	0.32
α_{22} = 3	**JMIV**	-0.02	93.5	0.37
	TSLM	-0.38	5.5	0.44
	TSLMM	-0.01	97.0	0.49
	TSIV	0.01	95.0	0.43
γ_{11} = 2	**JMIV**	0.01	94.0	0.52
	TSLM	-0.43	26.5	0.64
	TSLMM	-0.38	15.0	0.35
	TSIV	0.03	76.0	0.41
γ_{12} = -1	**JMIV**	0.00	95.5	0.86
	TSLM	0.38	62.5	0.94
	TSLMM	-0.41	31.0	0.62
	TSIV	0.04	88.5	0.86
γ_{22} = 2	**JMIV**	0.00	98.0	0.43
	TSLM	0.51	2.0	0.46
	TSLMM	0.62	0.0	0.32
	TSIV	-0.01	88.5	0.40

TABLE 3

Simulation I: bias, coverage, and 95% credible interval (or confidence interval) length across 200 simulation replicates. We compare our Bayesian joint model (JMIV) to the 1) simple two-stage linear regression (TSLM) 2) the two-stage linear mixed model-linear regression (TSLMM) and 3) the two-stage individual variances (TSIV) model. See Section 4 for details about the alternative methods.

For the TSLMM approach, the coverage and bias of the α parameters have significantly improved compared to the TSLM approach, likely due to the linear mixed model appropriately capturing the dependence between individuals' data points (i.e., appropriately capturing the measurement error in the mean parameters). However, TSLMM still has difficulty in recovering the coefficients of the variances and covariances, as can be seen by the poor coverage and high bias of these parameters. This makes sense since this framework assumes that individual random effects variability can be drawn from a population level variance-covariance matrix (not capturing measurement error in the variance parameters). This result suggests that if the individual variances and covariances do have an influential role in estimating the outcome, neither TSLMM nor TSLM will be able to recover the true values of these parameters. Interestingly, the TSLMM results also show an attenuation towards the null for the γ parameters, but not for the α parameters (although the bias is negligible). This indicates that the TSLMM alternative is able to better estimate the individual intercepts and slopes, but not the residual variability.

Compared to the TSLM and the TSLMM approaches, the TSIV approach has noticeably better coverage and lower bias of the γ parameters. However, compared to our proposed

JMIV, TSIV is still uniformly 'worse' across the three metrics. The bias of the three $\boldsymbol{\gamma}$ parameters is higher when compared to the bias produced by JMIV. Also, none of the $\boldsymbol{\gamma}$ parameters have higher than 90% coverage across the 200 replicates and the average length of the 95% credible intervals is higher than the 95% credible intervals from the JMIV approach. Across all of the simulation replicates, JMIV achieved greater than 90% coverage of the true parameters. JMIV also achieved low bias across the simulation replicates. We do note that the average 95% CrI interval lengths are larger for the $\mathbf{R}_i$ parameters than for the $\mathbf{D}_i$ parameters (see Supplementary Material S5). This is likely due to the higher uncertainty in estimating these correlation parameters, which has been captured appropriately. This higher uncertainty is also likely the same mechanism behind the larger average 95% CrI interval length of the γ_{12} parameter (corresponding to the covariance of the two trajectories). Overall, these results demonstrate that our model is able to successfully recover the data generating parameters while maintaining good coverage and low bias.

5.2. *Simulation 2: Three Biomarkers.* Here we again simulate $n_i = 6$ to 15 timepoints each for $N = 1{,}000$ individuals. The simulated longitudinal data is generated by $X_{itq} = b_{iq1} + b_{iq2}t + \epsilon_{iq}, q = 1, 2,$ where $\boldsymbol{B}_{i1} \sim \mathcal{N}_2(\boldsymbol{\beta}_1, \Sigma_1), \boldsymbol{B}_{i2} \sim \mathcal{N}_2(\boldsymbol{\beta}_2, \Sigma_2), \boldsymbol{B}_{i3} \sim \mathcal{N}_2(\boldsymbol{\beta}_3, \Sigma_3)$, and $\boldsymbol{\beta}_1 = (0, 2)^\mathsf{T}$, $\boldsymbol{\beta}_2 = (2, 1)^\mathsf{T}$, $\boldsymbol{\beta}_3 = (1, 1)^\mathsf{T}$,

$$\Sigma_1 = \begin{pmatrix} 1 & -0.05 \\ -0.05 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & -0.1 \\ -0.1 & 0.5 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 1 \end{pmatrix},$$

$$\log(d_{i1}) \sim \mathcal{N}(0, 0.75)/2, \log(d_{i2}) \sim \mathcal{N}(0.5, 0.5)/2, \log(d_{i3}) \sim \mathcal{N}(0, 1)/2.$$

We first generate the following values $(c_{12}+1)/2 \sim \mathsf{Beta}(1, 5)$, $(c_{13}+1)/2 \sim \mathsf{Beta}(1, 5)$, $(c_{23}+1)/2 \sim \mathsf{Beta}(2, 2)$ and use the approach for $Q = 3$ markers described in Section 3.3 to generate the individual correlation matrices, $\mathbf{R}_i$.

Truth	Model	Bias	Coverage (%)	Average Interval Length
α_{11} = -3	**JMIV**	0.01	96.5	0.49
	TSLM	0.83	0.0	0.65
	TSLMM	0.00	93.5	0.69
	TSIV	-0.01	96.0	0.64
α_{12} = -3	**JMIV**	-0.01	93.0	0.45
	TSLM	1.02	0.0	0.53
	TSLMM	-0.01	94.0	0.62
	TSIV	0.00	94.5	0.55
α_{13} = 3	**JMIV**	-0.01	96.5	0.50
	TSLM	-0.57	4.5	0.86
	TSLMM	0.03	93.0	0.74
	TSIV	-0.01	92.0	0.59
α_{21} = -3	**JMIV**	-0.01	95.0	0.43
	TSLM	1.20	0.0	0.60
	TSLMM	0.003	95.0	0.73
	TSIV	-0.01	96.0	0.81
α_{22} = 3	**JMIV**	-0.01	92.5	0.63
	TSLM	0.85	2.5	0.36
	TSLMM	0.02	93.5	0.90
	TSIV	0.01	93.5	0.65
α_{23} = 3	**JMIV**	-0.01	97.5	0.47
	TSLM	0.14	82.5	0.67
	TSLMM	-0.02	91.5	0.65
	TSIV	0.01	96.0	0.59

TABLE 4

Simulation II: bias, coverage, and 95% credible interval (or confidence interval) length across 200 simulation replicates for the $\boldsymbol{\alpha}$ parameters. We compare our Bayesian joint model (JMIV) to the 1) simple two stage linear regression (TSLM) 2) the two stage linear mixed model-linear regression (TSLMM) and 3) the Bayesian two stage model (TSIV).

Truth Parameter	Model	Bias	Coverage (%)	Average Interval Length
γ_{11} = 2	**JMIV**	-0.01	93.5	1.08
	TSLM	-0.11	82.5	0.67
	TSLMM	-0.44	43.0	0.80
	TSIV	-0.02	83.0	0.97
γ_{12} = -1	**JMIV**	-0.04	92.0	1.66
	TSLM	-0.11	9.5	1.70
	TSLMM	1.22	8.0	1.28
	TSIV	-0.08	91.5	1.77
γ_{22} = 2	**JMIV**	0.03	96.5	0.79
	TSLM	-0.76	2.5	0.71
	TSLMM	-1.24	0.0	0.59
	TSIV	-0.03	87.5	0.76
γ_{13} = -2	**JMIV**	0.03	94.5	1.91
	TSLM	1.57	17.5	2.12
	TSLMM	1.62	2.5	1.29
	TSIV	-0.06	82.0	1.78
γ_{23} = 2	**JMIV**	0.04	94.5	1.50
	TSLM	-0.98	30.5	1.56
	TSLMM	-0.92	14.5	1.03
	TSIV	-0.01	90.5	1.52
γ_{33} = 1	**JMIV**	0.01	94.0	0.54
	TSLM	0.15	83.5	0.73
	TSLMM	-0.002	69.0	0.37
	TSIV	-0.02	64.4	0.42

TABLE 5

Simulation II: bias, coverage, and 95% credible interval (or confidence interval) length across 200 simulation replicates for the $\boldsymbol{\gamma}$ parameters. We compare our Bayesian joint model (JMIV) to the 1) simple two stage linear regression (TSLM) 2) the two stage linear mixed model-linear regression (TSLMM) and 3) the Bayesian two stage model (TSIV).

For the outcome submodel, we set the true values of the regression coefficients as: $\boldsymbol{\alpha} = (-3, -3, 3, -3, 3, 3), \boldsymbol{\gamma} = (2, -1, 2, -2, 2, 1)$, where these values were again chosen so that the distribution of the outcome y_i would be similar to the distribution of the SWAN body mass outcomes. Lastly, we set $\sigma^2 = 1$ (the variance parameter for the outcome).

We present the results of this simulation study in Tables 4 and 5. We note that the proposed JMIV achieves above 90% coverage for both the mean ($\boldsymbol{\alpha}$) and variance-covariance ($\boldsymbol{\gamma}$) parameters, which the other models fail to do. With respect to the $\boldsymbol{\alpha}$ parameters, TSLMM and TSIV both perform well in terms of both coverage and bias. However, substantial differences in performance are present in the $\boldsymbol{\gamma}$ parameters. We see that that our model, JMIV, consistently has lower bias except in the case of γ_{23} where TSIV achieves lower bias and γ_{33} where TSLMM achieves lower bias. However, in both cases, JMIV outperforms the other models in terms of higher coverage (substantially higher coverage in the case of γ_{33}), indicating that JMIV is still a better model choice.

6. Hormone Trajectories and Changes in Body Mass Across the Menopausal Transition. We examine the joint association of E2 and FSH on rate of change in fat mass and waist circumference as women transition from the premenopause to the postmenopause using data from the Study of Women's Health Across the Nation (SWAN). Previous research with SWAN data has demonstrated that fat mass increases and lean mass decreases in a nonlinear fashion across the menopausal transition (Greendale et al., 2019). The rate of change in body fat mass and lean mass accelerates approximately two years prior to the final menstrual period (FMP) and persists until approximately 2 years after the FMP (Greendale 2013). Body weight and BMI, however, have a consistent positive linear relationship throughout the menopause transition, suggesting no unique menopausal effect on body weight or BMI despite the changes to a more adverse body composition profile (i.e., more fat mass and less lean mass). Given that increases in weight and fat mass in midlife contribute to women's

risk of chronic disease, improving understanding of the physiologic mechanisms that underlie these increases is important. Yet, the association of mean body size parameters with both mean E2 and mean FSH is complex, especially late in the menopausal transition and into the postmenopause, because fat is a significant source of estrogen and a known negative feedback regulator of FSH in the hypothalamus and pituitary. The role of fat in moderating a woman's endocrine profile may help explain why, in women who are obese, E2 is lower prior to menopause and higher postmenopause while mean FSH is much lower, compared to women who are not obese (Randolph et al., 2011). Since not all fat is metabolically equal and functionally varies by anatomic distribution, a model that can evaluate the contribution of individual variability in hormones would advance understanding of the complex relationship between body composition and reproductive hormones.

In the predictor submodels, we specified a linear mean trend consisting of an individual intercept and slope. We also explored higher order quadratic forms for the mean trend, but found that the quadratic terms (e.g. E2 Intercept^2) did not significantly predict either outcome of interest, likely due to the sample size of our datasets, as well as the limited individual-level information that remained after detrending the hormone population trends. In the outcome submodels, we used the correlation between E2 and FSH rather than the covariance as a variable of interest, since the correlation measure has a more straightforward interpretation and is normalized to the E2, FSH variances.

Figure 3 shows the estimated trajectories from the model for two women in our dataset. WWe do note that there is some undersmoothing regarding the observed residuals and the model-predicted mean trajectories. The interpretation of the individual coefficients becomes difficult, if not impossible, with a complex mean structure. Since our main goal was to relate the means, variances, and covariances of the hormone markers to the outcome while preserving interpretability in the scientific application, we chose to proceed with the linear intercept-slope mean model.

In the outcome regression model, we adjusted for the following covariates: fat mass body weight proportion (or waist circumference) at the 'first' visit, race/ethnicity (White, Black, Chinese and Japanese) and sports activity category. We included race/ethnicity in the models given previous research using SWAN data that found differences in body mass composition changes among ethnic groups during the menopausal transition (Greendale et al., 2019, 2021). The physical activity category is a measure of the individual level physical activity trajectories for each subject in the SWAN study, grouped into categories reflecting: (1) lowest, (2) increasing, (3) decreasing, (4) middle, and (5) highest physical activity during follow up. For a more detailed description, please refer to Pettee Gabriel et al. (2017). Tables 1 and 2 display the descriptive statistics for the individuals in our two analyses, including demographic and physical activity information. The following sections contain the results from applying our joint model to the SWAN datasets.

6.1. *Fat Mass Rate of Change.* Table 6 displays the results of the fat mass model. For ease of interpretation, the coefficients relating to the individual means (E2, FSH intercepts and E2, FSH slopes) and the individual variances (E2, FSH variances and E2, FSH correlation) have been multiplied by their respective sample standard deviation estimates. Table S3 in the Supplementary Material displays these SD estimates.

We found that the E2 intercept and the E2 and FSH slopes were all significantly associated with fat mass rate of change. A one standard deviation higher E2 intercept (compared to the population mean) was was associated with an average 0.11% increase in fat mass proportion per year. Since E2 tends to decline over the menopausal transition, we can interpret the E2 coefficient as follows: one standard deviation lower E2 slope than the population average was associated with a mean decrease of 0.09% in fat mass proportion per year. Conversely,

Variable	Post. Mean	2.5% CrI	97.5% CrI
E2 Intercept	**0.11**	**0.05**	**0.17**
FSH Intercept	0.02	-0.01	0.06
E2 Slope	**0.09**	**0.02**	**0.17**
FSH Slope	**-0.06**	**-0.11**	**-0.02**
E2 Var.	-0.03	-0.08	0.02
E2, FSH Cor.	0.01	-0.05	0.07
FSH Var.	-0.03	-0.08	0.01
Fat Mass Prop. (First Visit)	**-2.97**	**-3.44**	**-3.51**
Black	**-0.08**	**-0.16**	**-0.01**
Chinese	0.01	-0.08	0.11
Japanese	**-0.32**	**-0.41**	**-0.24**
Increasing Activity (Cat. 2)	0.02	-0.07	0.11
Decreasing Activity (Cat. 3)	0.07	-0.01	0.15
Middle Activity (Cat. 4)	0.01	-0.06	0.09
Highest Activity (Cat. 5)	-0.09	-0.17	0.00

TABLE 6

Estimated posterior means and 95% credible intervals for the fat mass rate of change model. The variables related to E2 and FSH (intercepts, slopes, and variances/correlation) have been standardized by the sample standard deviation of their posterior estimates. The presented values have been multiplied by 10^2.

higher increases in FSH were negatively associated with fat mass rate of change, with a one standard deviation increase in FSH (compared to the population average) being associated with -0.06% increase in fat mass proportion per year.

Fat mass proportion at first visit was negatively associated with fat mass rate of change (-2.97% decline per year). Black and Japanese women also had slower fat mass gains compared to white women on average (-0.08% increase per year and -0.32% per year, respectively).

6.2. *Waist Circumference Rate of Change.* Initially the Gaussian outcome assumption did not appear to be a good fit for the observed outcome. In particular, the residuals suggested overdispersed variances with a common mean, so we allowed the outcome to be modeled as a mixture of two Gaussian distributions with equal means but different variances:

$$
\begin{aligned}
&Y_i | z_i, \eta(\mathbf{B}_i, \mathbf{S}_i, \boldsymbol{W}_i), \sigma_1^2, \sigma_2^2 \sim \mathcal{N}(\eta_i, \sigma_{z_i}^2), \\
&z_i | \pi \sim \mathsf{Bernoulli}(\pi), \\
&\pi \sim \mathsf{Beta}(1/2, 1/2), \sigma_1 \sim \mathsf{half-Cauchy}(0, 2.5), \sigma_2 \sim \mathsf{half-Cauchy}(0, 5),
\end{aligned}
$$

where z_i is an unobserved indicator variable indicating membership in the first mixture component; $z_i = 1$ for the first component. Because the mean is equal across the mixture components, the interpretation of the regression parameters will be the same as for the fat mass models despite the additional variance parameter.

Table 7 displays the estimated coefficients for the waist circumference model. As in the fat mass model, the coefficients related to the individual means and variances have been adjusted by their respective sample standard deviations (also found in Table S3 in the Supplementary Material). A one-unit higher E2 intercept (above the population mean) was associated with an average 0.19 faster increase in waist circumference (cm/year). A one standard deviation lower E2 slope was also associated with slower declines in waist circumference per year (-0.26 cm/year). A higher individual FSH intercept was negatively associated with waist circumference rate of change; a one-unit higher starting FSH was associated with an average -0.07 cm/year decrease in waist circumference.

E2 variability was negatively associated with the outcome, meaning that women with a one standard deviation higher E2 variability had, on average, -0.11 cm/year decrease in waist circumference. Neither FSH variability nor E2, FSH correlation were significantly associated with changes in waist circumference.

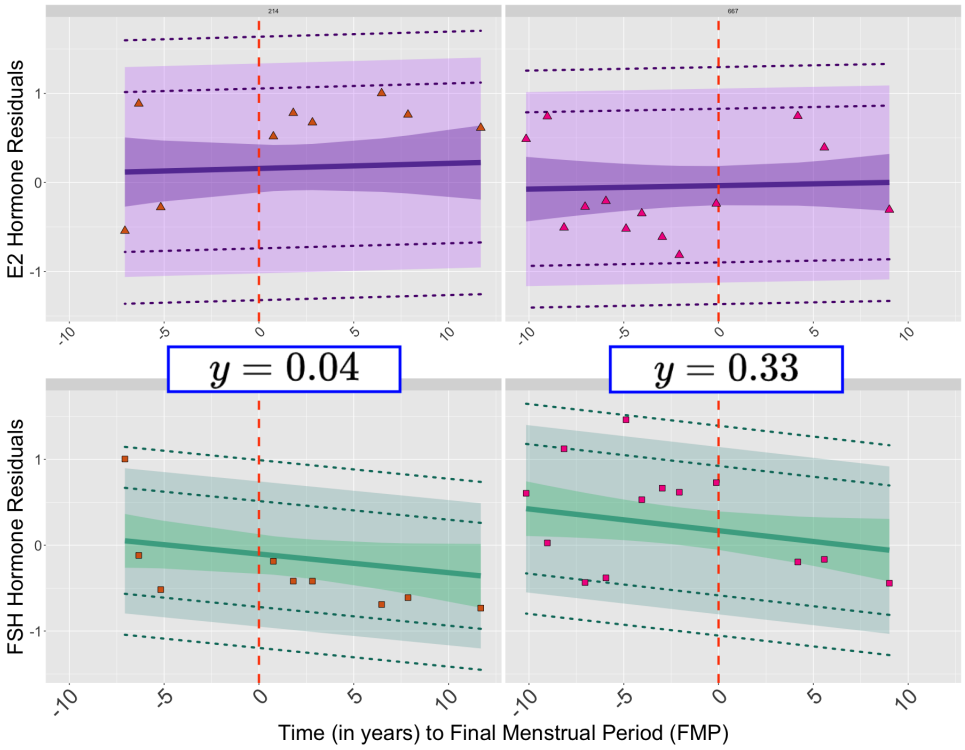

FIG 3. *Plots of estimated hormone trajectories for two individuals from the waist circumference rate of change model. The solid lines are the estimated individual mean trajectories, based on the posterior means of* $\boldsymbol{B_i}$*, i.e.* $\widehat{b}_{i0} + \widehat{b}_{i1}t$*. The darker inner intervals around the solid lines are* $+/-1.64 \times var(\widehat{b}_{i0} + \widehat{b}_{i1}t)$ *and the lighter band is* $+/-1.64 \times \widehat{\sigma}_{iq}$*, where* $\widehat{\sigma}_{iq}$ *is the square root of the estimated posterior mean of the individual level variance of hormone q. The dotted lines represent* $+/-1.64 \times \sigma_{iq5}$ *and* $+/-1.64 \times \sigma_{iq95}$ *where* $\sigma_{iq5}, \sigma_{iq95}$ *are the values of the 5th and 95th percentiles of the posterior samples of the individual variances for each hormone q. The triangles and squares are the observed E2 and FSH residuals, respectively. The observed individual waist circumference rates of change are shown in bordered boxes.*

Variable	Post. Mean	2.5% CrI	97.5% CrI
E2 Intercept	**0.19**	**0.11**	**0.29**
FSH Intercept	**-0.07**	**-0.13**	**-0.19**
E2 Slope	**0.26**	**0.16**	**0.37**
FSH Slope	**-0.17**	**-0.11**	**-4.35**
E2 Var.	**-0.11**	**-0.13**	**-0.03**
E2, FSH Cor.	-0.03	-0.12	0.06
FSH Var.	-0.05	-0.82	0.11
Waist Circum. (First visit)	**-0.02**	**-0.02**	**-0.01**
Black	-0.08	-0.20	0.03
Chinese	**-0.30**	**-0.45**	**-0.16**
Japanese	**-0.24**	**-0.38**	**-0.10**
Hispanic	-0.06	-0.33	0.21
Increasing Activity (Cat. 2)	0.00	-0.14	0.15
Decreasing Activity (Cat. 3)	-0.04	-0.16	0.08
Middle Activity (Cat. 4)	-0.09	-0.21	0.04
Highest Activity (Cat. 5)	-0.13	-0.27	0.02

TABLE 7

Estimated posterior means and 95% credible intervals for the waist circumference rate of change model. The variables related to E2 and FSH (intercepts, slopes, and variances/correlation) have been standardized by the sample standard deviation of their posterior estimates.

Unsurprisingly, waist circumference at first visit was negatively associated with waist circumference change, meaning that women with a higher starting waist circumference tended to have slower increases in waist circumference (-0.02 cm/year increase). Chinese and Japanese women also had slower increases compared to white women on average (-0.30 cm/year increase and -0.24 cm/year increase, respectively).

7. Discussion. We have presented a joint modeling approach for estimating individual-level mean and variance-covariance matrices based on longitudinal marker trajectories, which are then linked to a cross-sectional outcome. Simulations show that our model outperforms alternative approaches to this research problem. Our analysis of hormone trajectories data revealed E2 variability had a statistically significant association with waist circumference change, but not overall body mass composition, across the menopausal transition.

Our work is important for both methodological development of joint models and for women's health research. Our model estimates both mean longitudinal trends and the residual variability of these individual trajectories, and propagates the estimation uncertainty into the second submodel. This joint modeling is important for obtaining accurate estimates (in terms of low bias, higher coverage and shorter interval lengths) of how the individual-level parameters are linked to the outcome. Simulation results demonstrate that our model outperforms common two-stage approaches.

Substantively, our analyses are in line with the established literature on the associations between average hormone levels and fat mass and distribution changes during menopause. As noted above, the association of mean body mass with both mean E2 and mean FSH is complex, especially as women transition into the postmenopause, because adipose tissue is a significant source of estrogen and a known negative feedback regulator of FSH in the hypothalamus and pituitary. The known E2 results are echoed by our analyses, which showed a 1) positive relationship between increasing E2 and fat mass gains and 2) a positive relationship between increasing E2 and waist circumference gains and 3) a negative relationship between increasing FSH and both outcomes of interest. However, evidence suggests that increased FSH itself may directly influence adiposity by reductions in energy expenditure after menopause Sponton and Kajimura, 2017; Kohrt and Wierman, 2017; Liu et al., 2017. Thus, our findings of a negative relationship between increasing FSH and both outcomes of interest do not support this recent work. This may be due to the complex and complicated relationship of concurrent E2 changes during the menopausal transition.

However, the associations between individual-level variability and co-variability of two hormones (E2 and FSH) and changes in fat mass and waist circumference, a surrogate for fat distribution, had not been well explored. Previous analyses have either only evaluated mean associations of E2 and FSH on health outcomes, or separately analyzed the hormone's variability. Early models of hormone mean trajectories were far too crude to assess individual variability or associations with fat distribution. Our analyses revealed that individual E2 variability was highly predictive of waist circumference changes, but not overall fat mass changes. Since not all fat is metabolically equal and functionally varies by anatomic distribution, this finding could indicate that changes in fat distribution, in particular waist adiposity, are more driven by E2 hormonal variability during menopause, while other factors could be driving overall fat mass increases. Future analyses would be required to more fully investigate that hypothesis. Additionally, this joint analysis of E2 and FSH can serve as a basis for further investigation of how hormone variability and co-variability may affect other health outcomes. As mentioned above, joint estimation of longitudinal variables and scalar outcomes can be useful for investigating scientific questions in many areas. With longitudinal biomarker data becoming more readily available (e.g. from wearable devices), we need statistical methods for analyzing these types of data. Our proposed method addresses the gap in

methods by 1) providing a framework for jointly modeling longitudinal and cross-sectional data and 2) explicitly modeling individual-level variability in the longitudinal trajectories, which can improve understanding of the relationship between longitudinal predictors and health outcomes.

7.1. *Remark.* In our simulation studies and SWAN data analysis, we made the simplifying assumption to exclude covariates $\boldsymbol{W}_i$ in modeling the longitudinal markers (Equation 5). In mathematical terms, this means that the likelihood functions of $\boldsymbol{X}_{it}, \boldsymbol{B}_i, \boldsymbol{S}_i, \boldsymbol{W}_i$ are:

$$f(\boldsymbol{X}_{it} \mid \boldsymbol{B}_i, \boldsymbol{S}_i, \boldsymbol{W}_i) = f(\boldsymbol{X}_{it} \mid \boldsymbol{B}_i, \boldsymbol{S}_i),$$
$$f(\boldsymbol{B}_i, \boldsymbol{S}_i \mid \boldsymbol{W}_i) = f(\boldsymbol{B}_i, \boldsymbol{S}_i),$$

where f is a generic notation for the probability density function. For the scientific application, our main focus was to evaluate the overall marginal effects of the biomarker hormone means and variances on the body mass outcomes of interest. In particular, the effects of the variance and correlation parameters were of key interest, since the associations between individual E2 and FSH variabilities (and co-variability) and body mass changes had not been previously explored. The estimated coefficients for the $\boldsymbol{B}_i, \boldsymbol{S}_i$ described in Section 6 should be interpreted as marginal effects, rather than conditional on the other adjusted covariates. For this particular scientific application, we believed that this assumption resulted in a more straightforward interpretation of the mean, variance, and correlation parameters. The decision to include or exclude $\boldsymbol{W}_i$ in the longitudinal submodel should be made with the specific research application in mind and whether or not the simplifying assumption makes sense for the particular context.

7.2. *Future Work.* One extension of this work could be to model the individual variances as being functions of time, i.e. $\mathbf{S}_{it}$. E2 and FSH are known to be highly variable as women approach their final menstrual cycle, so estimating $\mathbf{S}_{it}$ may better capture such changes in the biomarker variances. To obtain these estimates, we would likely need a larger dataset (with both more individuals and timepoints) than is currently available with the SWAN study. Another methodological extension could be to extend this model to account for missingness in both the trajectory data and the outcome data. For this analysis, we removed the missing values in the hormone data and only analyzed individuals with observed body mass outcomes. Although less than 5% of the values in our dataset were missing, analyzing complete case data only could still result in slightly biased inference. In the SWAN dataset, individuals can be subject to intermittent missingness as well as dropout; these types of missing data patterns could be addressed in future work. It may be of interest in future applications to simultaneously model multiple cross-section individual outcomes, as our outcome submodel specification only considered univariate outcomes. Exploring Bayesian semi-parametric approaches to modeling the subject specific parameters, e.g. with a Dirichlet process prior on the unknown parameters, would be another methodological extension. In addition to the increased model flexibility, this could also allow for clustering of individuals with similar mean trajectories and/or residual variances and covariances. Finally, we note that increasing the number of longitudinal trajectories may result in a form of $\eta(\cdot)$ in the OR model that is complicated to estimate, since the number of variance-covariance parameters increases quadratically with the number of trajectories. Some type of dimension reduction procedure may be useful in these settings, although retaining interpretability may be challenging.

Acknowledgments. The Study of Women's Health Across the Nation (SWAN) has grant support from the National Institutes of Health (NIH), DHHS, through the National Institute on Aging (NIA), the National Institute of Nursing Research (NINR) and the NIH Office of Research on Women's Health (ORWH) (Grants U01NR004061; U01AG012505, U01AG012535, U01AG012531, U01AG012539, U01AG012546, U01AG012553, U01AG012554, U01AG012495, and U19AG063720). This work also was supported by National Institute on Aging Grant 1-R56-AG066693. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIA, NINR, ORWH or the NIH. This research also was supported in part through computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor.

Clinical Centers: *University of Michigan, Ann Arbor – Carrie Karvonen-Gutierrez, PI 2021-present, Siobán Harlow, PI 2011-2021, MaryFran Sowers, PI 1994-2011; Massachusetts General Hospital, Boston, MA – Sherri-Ann Burnett-Bowie, PI 2020 – Present; Joel Finkelstein, PI 1999 – 2020; Robert Neer, PI 1994 – 1999; Rush University, Rush University Medical Center, Chicago, IL – Imke Janssen, PI 2020 – Present; Howard Kravitz, PI 2009 – 2020; Lynda Powell, PI 1994 – 2009; University of California, Davis/Kaiser – Elaine Waetjen and Monique Hedderson, PIs 2020 – Present; Ellen Gold, PI 1994 - 2020; University of California, Los Angeles – Arun Karlamangla, PI 2020 – Present; Gail Greendale, PI 1994 - 2020; Albert Einstein College of Medicine, Bronx, NY – Carol Derby, PI 2011 – present, Rachel Wildman, PI 2010 – 2011; Nanette Santoro, PI 2004 – 2010; University of Medicine and Dentistry – New Jersey Medical School, Newark – Gerson Weiss, PI 1994 – 2004; and the University of Pittsburgh, Pittsburgh, PA – Rebecca Thurston, PI 2020 – Present; Karen Matthews, PI 1994 - 2020.*

NIH Program Office: *National Institute on Aging, Bethesda, MD – Rosaly Correa-de-Araujo 2020 - present; Chhanda Dutta 2016- present; Winifred Rossi 2012–2016; Sherry Sherman 1994 – 2012; Marcia Ory 1994 – 2001; National Institute of Nursing Research, Bethesda, MD – Program Officers.*

Central Laboratory: *University of Michigan, Ann Arbor – Daniel McConnell (Central Ligand Assay Satellite Services).*

Coordinating Center: *University of Pittsburgh, Pittsburgh, PA – Maria Mori Brooks, PI 2012 - present; Kim Sutton-Tyrrell, PI 2001 – 2012; New England Research Institutes, Watertown, MA - Sonja McKinlay, PI 1995 – 2001.*

Steering Committee: Susan Johnson, Current Chair, Chris Gallagher, Former Chair

We thank the study staff at each site and all the women who participated in SWAN.

We also thank the University of Michigan SWAN team for providing the datasets used in our analysis. We also thank the kind users from the Stan Forums for providing suggestions for approaching the 3-trajectory modeling setting. With regards to model implementation, we would like to thank Daniel Barker, Brock Palen, and Logan A. Walls for their assistance with writing efficient model code and programming the simulation replicates to run on the University of Michigan ARC-TS computing cluster.

Funding. This work was supported by National Institute on Aging Grant 1-R56-AG066693.

REFERENCES

BARTLETT, J. W. and KEOGH, R. H. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. *Statistical Methods in Medical Research* **27** 1695–1708. https://doi.org/10.1177/0962280216667764

U. S. CENSUS BUREAU 2017 National Population Projections Tables: Main Series. Section: Government.

BÜRKNER, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* **80** 1–28. https://doi.org/10.18637/jss.v080.i01

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76** 1–32. https://doi.org/10.18637/jss.v076.i01

CARR, M. C. (2003). The Emergence of the Metabolic Syndrome with Menopause. *The Journal of Clinical Endocrinology & Metabolism* **88** 2404–2411. https://doi.org/10.1210/jc.2003-030242

CARROLL, R. (2003). Variances Are Not Always Nuisance Parameters. *Biometrics* **59** 211–220. https://doi.org/10.1111/1541-0420.t01-1-00027

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. CRC Press LLC.

CHI, Y.-Y. and IBRAHIM, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62** 432–445. https://doi.org/10.1111/j.1541-0420.2005.00448.x

COLLELUORI, G., CHEN, R., NAPOLI, N., AGUIRRE, L. E., QUALLS, C., VILLAREAL, D. T. and ARMAMENTO-VILLAREAL, R. (2018). Fat Mass Follows a U-Shaped Distribution Based on Estradiol Levels in Postmenopausal Women. *Frontiers in Endocrinology* **9** 315. https://doi.org/10.3389/fendo.2018.00315

DARSINI, D., HAMIDAH, H., NOTOBROTO, H. B. and CAHYONO, E. A. (2020). Health risks associated with high waist circumference: A systematic review. *Journal of Public Health Research* **9** 1811. https://doi.org/10.4081/jphr.2020.1811

ELLIOTT, M. R., SAMMEL, M. D. and FAUL, J. (2012). Associations between Variability of Risk Factors and Health Outcomes in Longitudinal Studies. *Statistics in Medicine* **31** 2745–2756. https://doi.org/10.1002/sim.5370

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1** 515–534. https://doi.org/10.1214/06-BA117A

GHOSH, R. P., MALLICK, B. and POURAHMADI, M. (2021). Bayesian Estimation of Correlation Matrices of Longitudinal Data. *Bayesian Analysis* **16** 1039–1058. https://doi.org/10.1214/20-BA1237

GORDON, J. L., RUBINOW, D. R., EISENLOHR-MOUL, T. A., LESERMAN, J. and GIRDLER, S. S. (2016). Estradiol variability, stressful life events, and the emergence of depressive symptomatology during the menopausal transition. *Menopause (New York, N.Y.)* **23** 257–266. https://doi.org/10.1097/GME.0000000000000528

GOURLAY, M. L., SPECKER, B. L., LI, C., HAMMETT-STABLER, C. A., RENNER, J. B. and RUBIN, J. E. (2012). Follicle-stimulating hormone is independently associated with lean mass but not BMD in younger postmenopausal women. *Bone* **50** 311–316. https://doi.org/10.1016/j.bone.2011.11.001

GREENDALE, G. A., STERNFELD, B., HUANG, M., HAN, W., KARVONEN-GUTIERREZ, C., RUPPERT, K., CAULEY, J. A., FINKELSTEIN, J. S., JIANG, S.-F. and KARLAMANGLA, A. S. (2019). Changes in body composition and weight during the menopause transition. *JCI Insight* **4** e124865. https://doi.org/10.1172/jci.insight.124865

GREENDALE, G. A., HAN, W., FINKELSTEIN, J. S., BURNETT-BOWIE, S.-A. M., HUANG, M., MARTIN, D. and KARLAMANGLA, A. S. (2021). Changes in Regional Fat Distribution and Anthropometric Measures Across the Menopause Transition. *The Journal of Clinical Endocrinology and Metabolism* **106** 2520–2534. https://doi.org/10.1210/clinem/dgab389

GRILICHES, Z. and INTRILIGATOR, M. D. (1987). *Handbook of Econometrics* 25, 1465-1514. North Holland.

HARLOW, S. D., LIN, X. and HO, M. J. (2000). Analysis of menstrual diary data across the reproductive life span applicability of the bipartite model approach and the importance of within-woman variance. *Journal of Clinical Epidemiology* **53** 722–733. https://doi.org/10.1016/s0895-4356(99)00202-4

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480. https://doi.org/10.1093/biostatistics/1.4.465

HUANG, X., ELLIOTT, M. R. and HARLOW, S. D. (2014). Modeling Menstrual Cycle Length and Variability at the Approach of Menopause Using Hierarchical Change Point Models. *Journal of the Royal Statistical Society. Series C, Applied Statistics* **63** 445–466. https://doi.org/10.1111/rssc.12044

IBRAHIM, J. G., CHU, H. and CHEN, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology* **28** 2796–2801. https://doi.org/10.1200/JCO.2009.25.0654

JIANG, B., ELLIOTT, M. R., SAMMEL, M. D. and WANG, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics* **71** 487–497. https://doi.org/10.1111/biom.12284

KARVONEN-GUTIERREZ, C. and HARLOW, S. D. (2017). Menopause and Midlife Health Changes. In *Hazzard's Geriatric Medicine and Gerontology* 7 ed. (J. B. Halter, J. G. Ouslander, S. Studenski, K. P. High, S. Asthana, M. A. Supiano and C. Ritchie, eds.) McGraw-Hill Education.

KOHRT, W. M. and WIERMAN, M. E. (2017). Preventing Fat Gain by Blocking Follicle-Stimulating Hormone. *The New England Journal of Medicine* **377** 293–295. https://doi.org/10.1056/NEJMcibr1704542

LAWRENCE GOULD, A., BOYE, M. E., CROWTHER, M. J., IBRAHIM, J. G., QUARTEY, G., MICALLEF, S. and BOIS, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues; Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* **34** 2181–2195. https://doi.org/10.1002/sim.6141

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100** 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

LIU, P., JI, Y., YUEN, T., RENDINA-RUEDY, E., DEMAMBRO, V. E., DHAWAN, S., ABU-AMER, W., IZADMEHR, S., ZHOU, B., SHIN, A. C., LATIF, R., THANGESWARAN, P., GUPTA, A., LI, J., SHNAYDER, V., ROBINSON, S. T., YU, Y. E., ZHANG, X., YANG, F., LU, P., ZHOU, Y., ZHU, L.-L., OBERLIN, D. J., DAVIES, T. F., REAGAN, M. R., BROWN, A., KUMAR, T. R., EPSTEIN, S., IQBAL, J., AVADHANI, N. G., NEW, M. I., MOLINA, H., VAN KLINKEN, J. B., GUO, E. X., BUETTNER, C., HAIDER, S., BIAN, Z., SUN, L., ROSEN, C. J. and ZAIDI, M. (2017). Blocking FSH induces thermogenic adipose tissue and reduces body fat. *Nature* **546** 107-112. https://doi.org/10.1038/nature22342

LONG, J. D. and MILLS, J. A. (2018). Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Medical Research Methodology* **18** 138. https://doi.org/10.1186/s12874-018-0592-9

MOHAMMADALIZADEH CHARANDABI, S., REZAEI, N., HAKIMI, S., MONTAZERI, A., TAHERI, S., TAGHINEJAD, H. and SAYEHMIRI, K. (2015). Quality of Life of Postmenopausal Women and Their Spouses: A Community-Based Study. *Iranian Red Crescent Medical Journal* **17** e21599. https://doi.org/10.5812/ircmj.21599

OGBURN, E. L., RUDOLPH, K. E., MORELLO-FROSCH, R., KHAN, A. and CASEY, J. A. (2021). A Warning About Using Predicted Values From Regression Models for Epidemiologic Inquiry. *American Journal of Epidemiology* **190** 1142–1147. https://doi.org/10.1093/aje/kwaa282

PAPAGEORGIOU, G., MAUFF, K., TOMER, A. and RIZOPOULOS, D. (2019). An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes. *Annual Review of Statistics and Its Application* **6** 223-240. https://doi.org/10.1146/annurev-statistics-030718-105048

PARK, S. K., HARLOW, S. D., ZHENG, H., KARVONEN-GUTIERREZ, C., THURSTON, R. C., RUPPERT, K., JANSSEN, I. and RANDOLPH, J. F. (2017). Association between changes in oestradiol and follicle-stimulating hormone levels during the menopausal transition and risk of diabetes. *Diabetic medicine : a journal of the British Diabetic Association* **34** 531–538. https://doi.org/10.1111/dme.13301

PETTEE GABRIEL, K., STERNFELD, B., COLVIN, A., STEWART, A., STROTMEYER, E. S., CAULEY, J. A., DUGAN, S. and KARVONEN-GUTIERREZ, C. (2017). Physical activity trajectories during midlife and subsequent risk of physical functioning decline in late mid-life: The Study of Women's Health Across the Nation (SWAN). *Preventive Medicine* **105** 287–294. https://doi.org/10.1016/j.ypmed.2017.10.005

PROUST-LIMA, C., SÉNE, M., TAYLOR, J. M. G. and JACQMIN-GADDA, H. (2014). Joint latent class models for longitudinal and time-to-event data: a review. *Statistical Methods in Medical Research* **23** 74–90. https://doi.org/10.1177/0962280212445839

RANDOLPH, J. F. JR., SOWERS, M., BONDARENKO, I. V., HARLOW, S. D., LUBORSKY, J. L. and LITTLE, R. J. (2004). Change in Estradiol and Follicle-Stimulating Hormone across the Early Menopausal Transition: Effects of Ethnicity and Age. *The Journal of Clinical Endocrinology & Metabolism* **89** 1555–1561. https://doi.org/10.1210/jc.2003-031183

RANDOLPH, J. F., ZHENG, H., SOWERS, M. R., CRANDALL, C., CRAWFORD, S., GOLD, E. B. and VUGA, M. (2011). Change in Follicle-Stimulating Hormone and Estradiol Across the Menopausal Transition: Effect of Age at the Final Menstrual Period. **96** 746–754. https://doi.org/10.1210/jc.2010-1746

REES, M., BITZER, J., CANO, A., CEAUSU, I., CHEDRAUI, P., DURMUSOGLU, F., ERKKOLA, R., GEUKES, M., GODFREY, A., GOULIS, D. G., GRIFFITHS, A., HARDY, C., HICKEY, M., HIRSCHBERG, A. L., HUNTER, M., KIESEL, L., JACK, G., LOPES, P., MISHRA, G., OOSTERHOF, H., PINES, A., RIACH, K., SHUFELT, C., VAN TROTSENBURG, M., WEISS, R. and LAMBRINOUDAKI, I. (2021). Global consensus recommendations on menopause in the workplace: A European Menopause and Andropause Society (EMAS) position statement. *Maturitas* **151** 55–62. https://doi.org/10.1016/j.maturitas.2021.06.006

RICHARDSON, S. and GILKS, W. R. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. **138** 430–442. https://doi.org/10.1093/oxfordjournals.aje.a116875

ROBERT, C. P. and CASELLA, G. (2010). Monte Carlo Integration. In *Introducing Monte Carlo Methods with R*, (C. Robert and G. Casella, eds.). *Use R* 61–88. Springer.

ROSS, R., NEELAND, I. J., YAMASHITA, S., SHAI, I., SEIDELL, J., MAGNI, P., SANTOS, R. D., ARSENAULT, B., CUEVAS, A., HU, F. B., GRIFFIN, B. A., ZAMBON, A., BARTER, P., FRUCHART, J.-C., ECKEL, R. H., MATSUZAWA, Y. and DESPRÉS, J.-P. (2020). Waist circumference as a vital sign in clinical practice: a Consensus Statement from the IAS and ICCR Working Group on Visceral Obesity. *Nature Reviews Endocrinology* **16** 177–189. Number: 3 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41574-019-0310-7

SAMMEL, M., WANG, Y., RATCLIIE, S., FREEMAN, E. and PROPERT, K. (2001). Models for within-subject heterogeneity as predictors for disease. In *Proceedings of the Annual Meeting of the American Statistical Association*.

SOWERS, M., CRAWFORD, S. L., STERNFELD, B., MORGANSTEIN, D., GOLD, E. B., GREENDALE, G. A., EVANS, D., NEER, R., MATTHEWS, K., SHERMAN, S., LO, A., WEISS, G. and KELSEY, J. (2000). SWAN: A Multicenter, Multiethnic, Community-Based Cohort Study of Women and the Menopausal Transition. In *Menopause: Biology and Pathology* (R. A. Lobo, J. Kelsey and R. Marcus, eds.) 175–188. Academic Press. https://doi.org/10.1016/B978-012453790-3/50012-3

SOWERS, M., ZHENG, H., TOMEY, K., KARVONEN-GUTIERREZ, C., JANNAUSCH, M., LI, X., YOSEF, M. and SYMONS, J. (2007). 6-year changes in body composition in women at mid-life: ovarian and chronological aging. *The Journal of clinical endocrinology and metabolism* **92** 895–901. https://doi.org/10.1210/jc.2006-1393

SPONTON, C. H. and KAJIMURA, S. (2017). Burning Fat and Building Bone by FSH Blockade. *Cell Metabolism* **26** 285–287. https://doi.org/10.1016/j.cmet.2017.07.018

STEVENS, J., CAI, J., EVENSON, K. R. and THOMAS, R. (2002). Fitness and Fatness as Predictors of Mortality from All Causes and from Cardiovascular Disease in Men and Women in the Lipid Research Clinics Study. *American Journal of Epidemiology* **156** 832–841. https://doi.org/10.1093/aje/kwf114

STAN DEVELOPMENT TEAM (2020). RStan: the R interface to Stan. R package version 2.21.2.

STAN DEVELOPMENT TEAM (2023). 6.1 Bayesian measurement error model | Stan User's Guide.

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis* **16** 667–718. Publisher: International Society for Bayesian Analysis. https://doi.org/10.1214/20-BA1221

WANG, J., LUO, S. and LI, L. (2017). Dynamic prediction for multiple repeated measures and event time data: An application to Parkinson's disease. *The Annals of Applied Statistics* **11** 1787–1809. https://doi.org/10.1214/17-AOAS1059

WANG, S., MCCORMICK, T. H. and LEEK, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences* **117** 30266–30275.

YOUNG, H. A. and BENTON, D. (2018). Heart-rate variability: a biomarker to study the influence of nutrition on physiological and psychological health? *Behavioural Pharmacology* **29** 140–151.

ZAIDI, M., LIZNEVA, D., KIM, S.-M., SUN, L., IQBAL, J., NEW, M. I., ROSEN, C. J. and YUEN, T. (2018). FSH, Bone Mass, Body Fat, and Biological Aging. *Endocrinology* **159** 3503–3514. https://doi.org/10.1210/en.2018-00601