Supporting Information
for
Multi-wave Validation Sampling for Error-prone Electronic Health Records
by

Bryan E. Shepherd, Kyunghee Han, Tong Chen, Aihua Bian, Shannon Pugh, Stephany N. Duda, Thomas Lumley, William J. Heerman, and Pamela A. Shaw

Web Appendix A

More Details on Phase 1 EHR Data

We received data for 20,684 mothers and 25,284 linked children. For mothers who delivered more than one child in separate pregnancies, we selected the first delivered child; in the case of multiple births from a single pregnancy, we randomly picked one child for inclusion. A small number of mothers ($n = 38$) with weight exceeding 180 kg (400 lbs) or whose weight was reported in the EHR to have changed more than 70 kg (150 lbs) during pregnancy were excluded.

Children's weight and height measurements during their first 6 years of life were cleaned using a validated algorithm developed by Daymon et al. (2017). A total of 12.3% (27,934 out of 226,272) of heights and 14.7% (85,919 out of 583,489) of weights were excluded using Daymon's method. Children's body mass index (BMI) was computed using heights and weights measured on the same day. If there were no same day measurements, then we used the nearest height measurement within ±3, ±7, ±14, and ±30 days for weights measured when children's ages were $<$ 90 days, 90-119 days, 120-729 days, and $\geq$ 730 days, respectively; 35% of children's heights were imputed in this manner. A total of 1.5% (2,946 out of 198,338 of heights and 39.5% (196,747 out of 497,570) of weights were excluded because of no corresponding weight/height measurement. Children's BMI percentile was calculated using the R package `childsds` (Vogel, 2019). Maternal BMI was computed using each mother's median height; measurements before the age of 15 years and extreme values $\leq$ 50 cm or $\geq$ 200 cm were excluded.

If the EHR indicated the mother had any smoking history prior to delivery, she was categorized as an ever smoker, otherwise, as never smoker.

Web Appendix B

FPCA Details

In this section we describe the FPCA procedure used to estimate maternal weight gain during pregnancy. Let $W_1(t), \ldots, W_N(t)$ denote a random sample of women's weights $W(t)$ at time t on a common domain $\mathcal{T} = [-365, 272]$ days, where $t = 0$ represents the date of conception and $t = 273$ is the date of birth. We assume measurements are independent between subjects and that $W(t)$ is smooth over $\mathcal{T}$. It follows from the Karhunen-Loève expansion that time-varying variations can be decomposed into linear combinations of eigenfunctions, the FPCA, such that

$$W_i(t) = \mu(t) + \sum_{k \geq 1} \xi_{ik} \phi_k(t), \tag{1}$$

where $\mu(t) = \mathrm{E}W_1(t)$ and ξ_{ik} are uncorrelated mean zero random variables with variance λ_k satisfying $\lambda_k \geq \lambda_{k+1}$ for any $k \geq 1$ (Ramsay and Silverman, 2007).

In our study, weights are measured at different time points such that the i-th mother has $\{W_i(t_{i1}), \ldots, W_i(t_{im_i})\}$ at time points $t_{i1} < \cdots < t_{im_i}$, where the number of measurements m_i also varies between mothers. In addition, we allow observed weight measurements to be contaminated by additive measurement errors $\widetilde{W}_{ij} = W_i(t_{ij}) + \epsilon_{ij}$, where ϵ_{ij} is an independent Gaussian error with mean zero, and $\widetilde{W}_{ij}$ is the error-prone weight phase 1 record of the i-th mother measured at the gestational age t_{ij}.

Yao et al. (2005) proposed the principal components analysis through conditional expectation (PACE) such that the best linear estimate of the FPC score ξ_{ik} is given by

$$\hat{\xi}_{ik} = \hat{\lambda}_k \widehat{\boldsymbol{\phi}}_{ik}^{\top} \widetilde{\Sigma}_i^{-1} (\widetilde{\mathbf{W}}_i - \widehat{\boldsymbol{\mu}}_i), \tag{2}$$

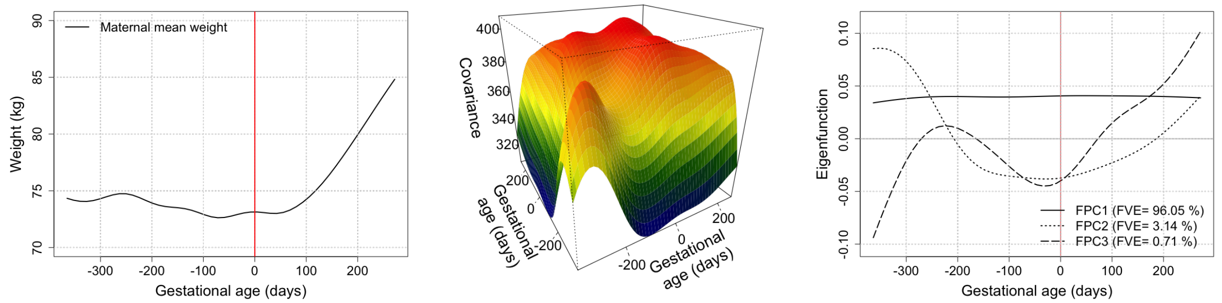
where $\widetilde{\mathbf{W}}_i = (\widetilde{W}_{i1}, \ldots, \widetilde{W}_{im_i})^{\top}$ are m_i-longitudinal observations, $\widehat{\boldsymbol{\mu}}_i = \{\hat{\mu}(t_{i1}), \ldots, \hat{\mu}(t_{im_i})\}^{\top}$

are estimates of $\mathrm{E}\widetilde{\mathbf{W}}_i = \{\mu(t_{i1}), \ldots, \mu(t_{im_i})\}^\top$, and $\widetilde{\Sigma}_i$ is the $m_i \times m_i$ variance-covariance matrix estimate of $\widetilde{\mathbf{W}}_i$. Here, $\hat{\lambda}_k$ and $\widehat{\boldsymbol{\phi}}_{ik} = \{\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{im_i})\}^\top$ are estimates of the eigenvalue λ_k and the evaluation of eigenfunction $\phi_k(t)$, respectively, where the pair $\{\lambda_k, \phi_k(t)\}$ is defined as the solution of the functional eigenequations given by Yao et al. (2005). We approximate the functional representation of the true time-varying trait $W_i(t)$ in (1) with the first leading K components of FPC scores $\hat{\xi}_{ik}$ and eigenfunctions $\hat{\phi}_k(t)$ as

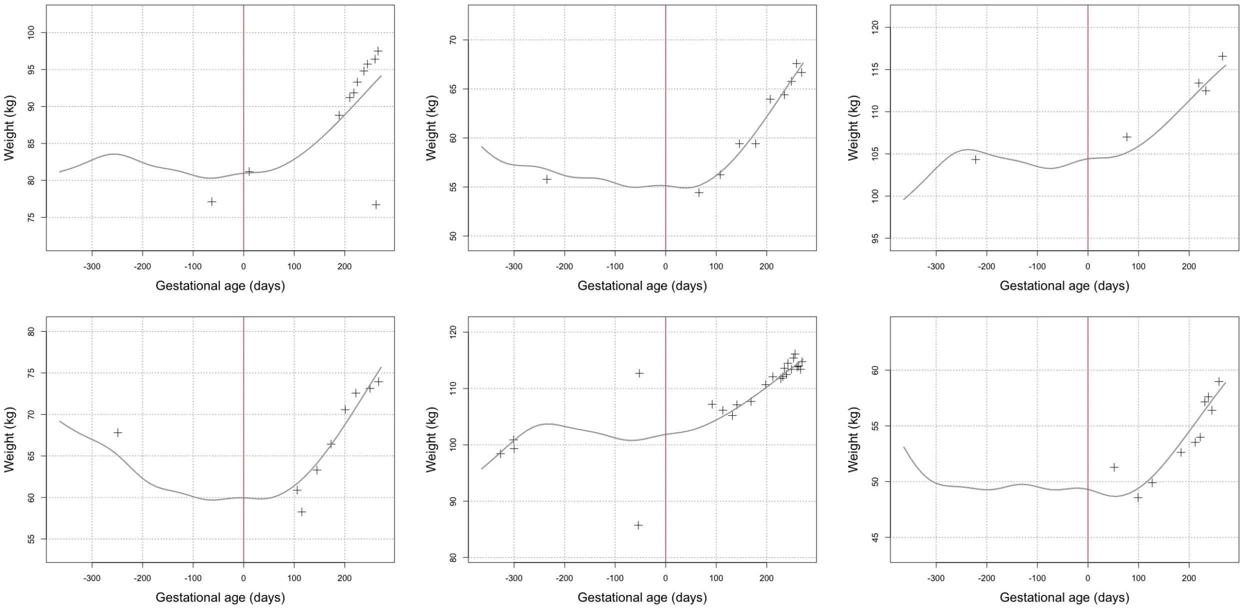
$$\widehat{W}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(t). \tag{3}$$

We refer to Yao et al. (2005) for technical details and confidence band estimation, and Ramsay and Silverman (2007) for an overview of functional data analysis. We used the R package `fdapace` (Carroll et al., 2021) for numerical implementation.

The FPCA results applied to the phase 1 data $\mathcal{W}_1$ suggested that mothers' weight trajectories can be well approximated using (3) with $K = 3$; the first three eigenfunctions explained 99.9% of the variance. Web Figure 1 shows the estimated mean function $\hat{\mu}(t)$, the estimated covariance function, and the first three eigenfunctions. Web Figure 2 depicts weight trajectories of six mothers constructed by the FPCA with the phase 1 data. The phase 1 weight at conception for mother i was estimated as $\widehat{W}_i(0)$, and the phase 1 exposure of interest, the maternal weight gain per week during pregnancy was given by $X_i^* = \left\{\widehat{W}_i(272) - \widehat{W}_i(0)\right\}/(273/7)$.



Web Figure 1: Estimation results of FPCA with the phase 1 data with mean function (left), covariance function (middle), and the first three eigenfunctions which explain 99.9% of functional variations in the data (right).



Web Figure 2: Weight trajectory estimates for six randomly chosen mothers.

Web Appendix C

More Details on Phase 2 Data Validation

The research nurse entered validation data into two spreadsheets and an electronic case report form using the Research Electronic Data Capture (REDCap) software (Harris et al., 2009). The two spreadsheets were used to reduce the data entry burden and number of button clicks to record audit findings for repeated values. The first spreadsheet contained maternal weights extracted from the EHR and the second spreadsheet contained children's heights and weights. All other phase 2 data were entered into the REDCap forms. We initially performed a pilot validation of 12 mother-child dyad records to refine our procedures and forms; validated data from the pilot was excluded from analyses. In the pilot validation, we realized that manual entry of dozens of weights per mother-child dyad would be extremely time-consuming, yield a small proportion with errors that needed to be fixed, and could result in validation data entry errors. Therefore, for our phase 2 sample, the research nurse only validated the following phase 1 measurements: children's heights/weights closest to their birthdays, children's heights/weights that led to a first diagnosis of obesity, maternal weight closest to but prior to delivery, maternal weight closest to but prior to 272 days before delivery, and any maternal weights flagged as potential outliers due to being outside the 95% confidence bands for the FPCA-predicted trajectories.

Web Appendix D

More Details on Multi-wave Sampling for Asthma Endpoint

Of the 750 records already validated for the obesity study, 582 met inclusion criteria for the asthma study. Our strategy was to 1) to use this already collected phase 2 data to build an imputation model for the validated data, 2) to impute "validated data" from that model for all mother-child records that had not been validated, 3) to fit a working analysis model to the complete data from which the influence function for the maternal weight gain log odds ratio was obtained, 4) to repeat this across multiple imputations to obtain the average influence function per mother-child dyad, and 5) to perform Neyman allocation based on these estimated average influence functions, refining strata so the allocation was approximately balanced across strata.

With regards to 1)-2), the validated estimated gestational age was first imputed using the R package `mice`; from this, the estimated maternal weight gain during pregnancy and BMI at conception were obtained from the FPCA. Then maternal asthma and child asthma were imputed using logistic regression models.

With regards to 3), our working outcome model was a logistic regression model with the outcome asthma (yes/no) based on the validated/imputed data; the exposure variable was the validated/imputed maternal weight change; validated/imputed covariates BMI at conception, estimated gestational age, and maternal asthma; and unvalidated covariates maternal race, maternal ethnicity, cesarean section, maternal age at delivery, and child sex.

We performed a total of 100 imputation replications, and computed the average influence functions per mother-child dyad across these imputation replications.

Web Appendix E

Additional Analysis of Obesity Endpoint Permitting Non-Linear Association

We performed an additional set of analyses that allowed for a non-linear relationship between maternal weight gain during pregnancy and childhood obesity. Although our primary, a priori specified analysis assumed a linear relationship, there is some thought that this relationship could be non-linear. Cox models which permitted this potential non-linear relationship were fit to both the error-prone phase 1 data and the validated phase 2 data. In both models, weight change during pregnancy was expanded using natural splines with two knots (same knot locations in both models corresponding to the 1/3 and 2/3 quantiles in the phase 1 data). All other covariates were included in these models as described in the main text. The model fit to the phase 2 data used generalized raking with inverse probability weights calibrated using estimates of the influence function for the three spline terms plugging in the phase 1 data (i.e., raking on the naive influence functions) and strata. Non-linearity of the associations were assessed using likelihood ratio tests.

Using the error-prone phase 1 data alone, there was little evidence of a non-linear association between maternal gain during pregnancy and the hazard of childhood obesity (p=0.87). However, our generalized raking estimator using the validated data suggested that there was a non-linear association (p=0.007). Web Table 1 shows estimates of the adjusted hazard ratio comparing specific values of maternal weight gain during pregnancy to 0.2 kg/wk. For context, the median maternal weight gain per week during pregnancy was 0.28 kg/wk (interquartile range 0.24, 0.35). The model suggested that after holding all other variables constant, a child from a woman who gained 0.6 kg/wk during pregnancy had a hazard of developing obesity that was approximately 66% higher than a woman who gained 0.2 kg/wk.

Web Table 1: Adjusted hazard ratios for childhood obesity based on maternal weight gain per week during pregnancy.

	Hazard Ratio	95% Confidence Interval
Average maternal weight gain per week during pregnancy (kg/wk)		
0	1.12	0.88, 1.43
0.1	1.02	0.93, 1.12
0.2 (reference)	1	
0.3	1.06	0.99, 1.14
0.4	1.20	1.03, 1.39
0.5	1.39	1.14, 1.70
0.6	1.66	1.32, 2.09

Analysis Code

Our analysis code can be found in the Biometrics website on Wiley Online Library. It is also posted at https://biostat.app.vumc.org/ArchivedAnalyses. We are unable to share our data.

References

Carroll, C., Gajardo, A., Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Müller, H.-G., and Wang, J.-L. (2021). *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.5.6.

Daymon, C., Ross, M. E., Russell Localio, A., Fiks, A. G., Wasserman, R. C., and Grundmeier, R. W. (2017). Automated identification of implausible values in growth data from pediatric electronic health records. *Journal of the American Medical Informatics Association*, 24:1080–1087.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (redcap)– a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42:377–381.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Vogel, M. (2019). *childsds: Data and Methods Around Reference Values in Pediatrics*. R package version 0.7.4.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.