Web-based Supplementary Materials for "Modeling longitudinal data with a random change point and no time-zero: applications to inference and prediction of the labor curve"

by Alexander C. McLain*[1] and Paul S. Albert[2]

[1]Department of Epidemiology and Biostatistics,
University of South Carolina,
800 Sumter Street, Columbia, SC 29208, USA

[2] Division of Intramural Population Health Research,
Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS,
6100 Executive Blvd Room 7B05, Rockville, MD 20852, U.S.A

Web Appendix A: Longitudinal Model

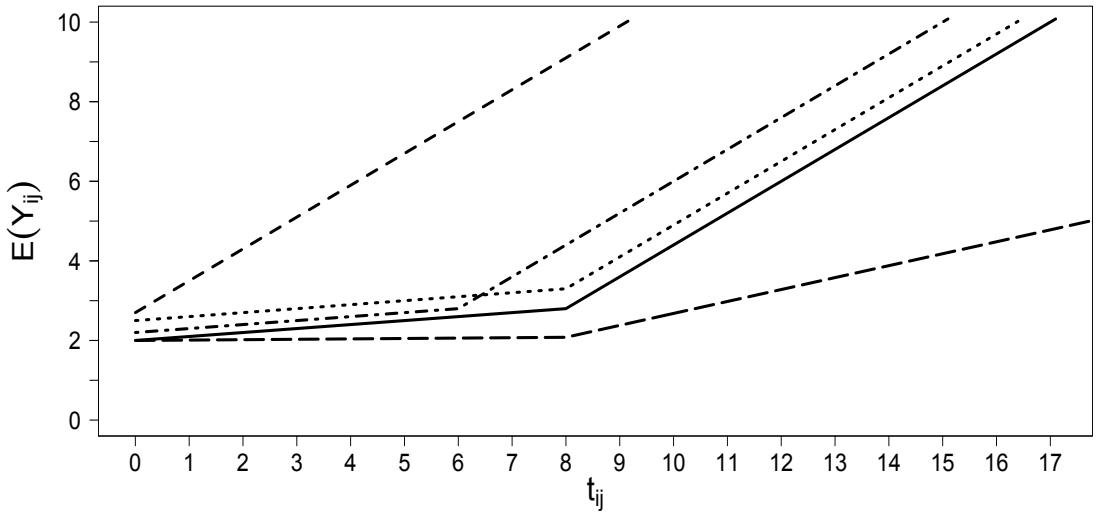
In this Section, we present some additional technical details on the longitudinal change point model. Figure 1 demonstrates some curves generated from the expectation of equation (1) in the main text. This expectation can be represented as

$$E\{Y_i(s_{ij})\} = \begin{cases} K + (\beta_1 + b_{i1})s_{ij} & \text{if } s_{ij} < c + b_{ic} \text{ and } c + b_{ic} \geq 0 \\ K + (\beta_1 + b_{i1})(c + b_{ic}) + (\beta_2 + b_{i2})\{s_{ij} - (c + b_{ic})\} & \text{if } s_{ij} \geq c + b_{ic} \text{ and } c + b_{ic} \geq 0 \\ K + (\beta_1 + b_{i1})\{s_{ij} - (c + b_{ic})\} + (\beta_2 + b_{i2})(c + b_{ic}) & \text{if } s_{ij} < c + b_{ic} \text{ and } c + b_{ic} < 0 \\ K + (\beta_2 + b_{i2})s_{ij} & \text{if } s_{ij} \geq c + b_{ic} \text{ and } c + b_{ic} < 0, \end{cases} \quad (1)$$

where $s_{ij} = t_{ij} - (\Delta + b_{i\Delta})$. The figure shows the expectation of the longitudinal model by t_{ij} (time since entry into the hospital) for different values of $\boldsymbol{b}_i$, where $\beta_1 = 0.1$, $\beta_2 = 0.8$, $\Delta = 0$, and $c = 8$.

Figure 1 demonstrates a few aspects of the model. First, we can see that the t at which the change point occurs decreases as $b_{i\Delta}$ or b_{ic} decreases (holding everything else constant). Second, for $\boldsymbol{b}_i = (0, 0, -5, 5)$ (dotted) versus $\boldsymbol{b}_i = \boldsymbol{0}$ (solid) the $E(Y_{ij})$ is increased by a constant. In general, if $\boldsymbol{b}_i = \boldsymbol{0}$ and $\boldsymbol{b}_i^* =$

*Corresponding Author, email:mclaina@mailbox.sc.edu



Web Figure 1: Realizations of $E(Y_{ij})$ versus t_{ij} with $K = 2$ by $\boldsymbol{b}_i = (b_{i1}, b_{i2}, b_{i\Delta}, b_{ic})$: $\boldsymbol{b}_i = \mathbf{0}$ (black solid), $\boldsymbol{b}_i = (0, 0, -2, 0)$ (dot–dash), $\boldsymbol{b}_i = (0, 0, 0, -9)$ (short dash), $\boldsymbol{b}_i = (0, 0, -5, 5)$ (dotted), and $\boldsymbol{b}_i = (-0.09, -0.5, 0, 0)$ (long dash), where $\beta_1 = 0.1$, $\beta_2 = 0.8$, $\Delta = 0$, and $c = 8$.

$(0, 0, -B, B)$ with $B > 0$ then $E\{Y_i(s)|\boldsymbol{b}_i\} + \beta_1 B = E\{Y_i(s)|\boldsymbol{b}_i^*\}$ if $c \geq 0$ and $E\{Y_i(s)|\boldsymbol{b}_i\} + \beta_2 B = E\{Y_i(s)|\boldsymbol{b}_i^*\}$ if $c < 0$, for all t. Thus, the model allows for varying mean cervical dilation at hospital entrance, reflecting entrance at different stages of labor. Third, the change point can occur before the women gets to the hospital (i.e., when $t < 0$). Fourth, the flexibility in the pre- and post-change point slopes allows for a variety of shapes that fit the data demonstrated in Figure 1 on the main text.

Web Appendix B: Adaptive Multivariate Rejection Sampling

A critical step in the MC algorithm presented in Section 4 of the main text, is the ability to draw random samples from the posterior density of $\boldsymbol{b}_i$. In this Section, we detail an adaptive multivariate rejection sampling algorithm that will generate a random sample of size L from the posterior density

$$h(\boldsymbol{b}_i|\mathcal{D}_{ik}, \boldsymbol{\theta}) \propto \prod_{j=1}^{J_i} f(Y_{ij}|\boldsymbol{b}_i, K, \boldsymbol{\theta}) g(\boldsymbol{b}_i|\Sigma) = L(\boldsymbol{b}_i|\mathcal{D}_{ik}, \boldsymbol{\theta}).$$

Multivariate rejection sampling is an algorithm that can generate random samples from a distribution without having to calculate the constant of proportionality. Given a candidate distribution h_c a random sample of size L from $h(\cdot|\mathcal{D}_{ik}, \boldsymbol{\theta})$ can be generated by iterating between the following three steps:

(a) draw $\boldsymbol{b}^c$ from a candidate distribution h_c, and w from a Uniform$(0,1)$ distribution,

(b) calculate $r = \tau_i^{-1} L(\boldsymbol{b}^c|\mathcal{D}_{ik}, \boldsymbol{\theta})/h_c(\boldsymbol{b}^c)$, where τ_i is such that $r \leq 1$ for all $\boldsymbol{b}^c$,

(c) if $w \leq r$ then accept $\boldsymbol{b}^c$, if $w > r$ then reject $\boldsymbol{b}^c$.

Steps (a)-(c) are iterated between until a sample of size L is obtained.

A standard choice for the candidate distribution is the estimated prior distribution of $\boldsymbol{b}_i$. However, this choice can be centered far away from the true posterior distribution with much larger variability resulting in a low acceptance probability. Let $h_{c3}(\cdot|\boldsymbol{\alpha}, \Psi)$ denote a multivariate t-distribution with non-centrality parameter $\boldsymbol{\alpha}$, covariance matrix Ψ, and degrees of freedom 3. For subject i in the mth iteration, the posterior distribution of $\boldsymbol{b}_i$ is approximated by h_{c3} setting $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_i$ and $\Psi = \hat{\Psi}_i$ where

$$\hat{\boldsymbol{\alpha}}_i = \text{argmax}_{\boldsymbol{b}_i} \left[\log\{L(\boldsymbol{b}_i|\mathcal{D}_{ik}, \boldsymbol{\theta}^m)\} \right], \tag{2}$$

the posterior mode of $h(\boldsymbol{b}_i|\mathcal{D}_i, \hat{\boldsymbol{\theta}}^m)$ and

$$\hat{\Psi}_i = \left[\frac{\partial^2}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i'} \log\{L(\boldsymbol{b}_i|\mathcal{D}_{ik}, \boldsymbol{\theta}^m)\} \right] \Big|_{\boldsymbol{b}_i = \hat{\boldsymbol{\alpha}}}, \tag{3}$$

the hessian matrix evaluated at $\hat{\boldsymbol{\alpha}}_i$. Given $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\Psi}_i$, τ_i in step (b) for a given value of $\hat{\boldsymbol{\theta}}^m$ is given by

$$\tau_i = \max_{\boldsymbol{b}} \left\{ \frac{L(\boldsymbol{b}|\mathcal{D}_{ik}, \boldsymbol{\theta}^m)}{h_{c3}(\boldsymbol{b}|\hat{\boldsymbol{\alpha}}_i, \hat{\Psi}_i)} \right\}. \tag{4}$$

Here $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\Psi}_i$ are estimated using the observed data $\mathcal{D}_{ik}$ and the current parameter vector $\hat{\boldsymbol{\theta}}^m$. As discussed in Section 3 of the main text, we use $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\Psi}_i$ in AGQ to center and scale the standard Gaussian quadrature nodes, respectively. For this reason they are a natural choice to improve the precision of the candidate distribution. By utilizing a t-distribution with 3 degrees of freedom we allow for extra room in the tails, which stabilizes the maximization in (4). The approximated h method will require extra optimization versus the standard method of setting the candidate distribution to the prior distribution. We found that approximated h method results in markedly higher acceptance probabilities than the standard method, which lead to substantially lower overall computation times.

Web Table 1: Summary simulation studies using linear covariate associations with $\boldsymbol{\beta}$, Δ, and c. The simulations were ran for $1,000$ iterations with $n = 500$. Displayed is the true parameter value ($\boldsymbol{\theta}$), the average of the estimated parameters (MEAN), and the Monte Carlo standard error (MCSE).

	$\boldsymbol{\theta}$	MEAN	MCSE		$\boldsymbol{\theta}$	MEAN	MCSE		$\boldsymbol{\theta}$	MEAN	MCSE
β_1	0.25	0.244	0.013	β_{11}	-0.10	-0.099	0.014	σ	0.25	0.255	0.009
β_2	1.50	1.503	0.018	β_{21}	0.20	0.199	0.018	$\rho_{b_1b_2}$	0.10	0.099	0.014
Δ	-0.75	-0.752	0.019	Δ_1	0.00	0.000	0.021	$\rho_{b_1\Delta}$	0.20	0.167	0.079
c	5.25	5.250	0.021	c_1	-0.50	-0.501	0.019	ρ_{b_1c}	-0.10	-0.138	0.090
σ_Δ	1.25	1.248	0.021	β_{12}	0.10	0.097	0.013	$\rho_{b_2\Delta}$	0.10	0.072	0.099
σ_c	1.24	1.241	0.021	β_{22}	-0.20	-0.198	0.015	ρ_{b_2c}	-0.20	-0.222	0.081
σ_{b_1}	0.24	0.238	0.016	Δ_2	0.50	0.499	0.017	$\rho_{c\Delta}$	0.30	0.254	0.095
σ_{b_2}	0.23	0.234	0.021	c_2	1.00	1.000	0.017				

Web Appendix C: Covariate Adjusted Simulation Study

In this Section, we present results from a simulation study that considered covariate effects on the parameters $\boldsymbol{\beta}$, Δ, and c. The data were generated using the model given in equation (1) of the main text, where $\boldsymbol{b}_i$ was generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. The diagonal elements of Σ are denoted by $(\sigma^2_{b_1}, \sigma^2_{b_2}, \sigma^2_\Delta, \sigma^2_c)$, with correlations $(\rho_{b_1b_2}, \rho_{b_1\Delta}, \rho_{b_1c}, \rho_{b_2\Delta}, \rho_{b_2c}, \rho_{\Delta c})$. The number of examination times, J_i, were generated via $J_i - 4 \sim$Poisson(5) so that there was an average of 9 examination times with a minimum of 4. The examination times were uniformly distributed over the interval $(0, 10)$. We induced linear covariate associations with $Z_1 \sim^{iid} N(0,1)$ and $Z_2 \sim^{iid}$Bernoulli(1/2), where $\beta_{1i} = \beta_1 + \beta_{11}Z_1 + \beta_{12}Z_2 + b_{i1}$, $\beta_{2i} = \beta_2 + \beta_{21}Z_1 + \beta_{22}Z_2 + b_{i2}$, $\Delta_i = \Delta + \Delta_1 Z_1 + \Delta_2 Z_2 + b_{i\Delta}$, and $c_i = c + c_1 Z_1 + c_2 Z_2 + b_{ic}$. The parameter values are given in Table 1 where $K = 2$.

For each of the 1000 simulated samples the procedure in Section 3 of the main text was implemented with 10 quadrature nodes. In Table 1 we present the results of the covariate adjusted simulation. For most of the parameters, the average estimated value demonstrated little bias. Similar to the covariate unadjusted simulation, some bias was apparent for the correlation parameters. For the parameters relating to the covariates ($\beta_{1j}, \beta_{2j}, \Delta_j, c_j$, for $j = 1, 2$) there was minimal bias. This demonstrates that the proposed method can be used to estimate multiple parameter associations with covariates that are binary or continuous.

Web Appendix D: Additional Data Analysis Results

In this section, we present some additional details of the data analysis presented in Section 6 of the main text. Specifically, we present the parameter estimates, standard errors and 95% confidence intervals for the

Web Table 2: The estimated standard deviation and correlation coefficients of the random effects for the covariate adjusted model (left), along with the normal (middle) and obese (right) BMI categories.

	Covariate adjusted			BMI 18.5 to 25kg/m^2			BMI 30 to 55kg/m^2		
	EST	SD	95% CI	EST	SD	95% CI	EST	SD	95% CI
σ_Δ	1.27	0.08	(1.11, 1.43)	2.23	0.04	(2.15, 2.30)	1.30	0.04	(1.23, 1.37)
σ_c	2.54	0.09	(2.37, 2.71)	2.87	0.05	(2.78, 2.97)	2.58	0.05	(2.47, 2.69)
σ_{b_1}	0.53	0.05	(0.43, 0.63)	0.22	0.02	(0.19, 0.25)	0.48	0.02	(0.44, 0.53)
σ_{b_2}	1.07	0.08	(0.91, 1.23)	0.59	0.02	(0.55, 0.63)	0.94	0.03	(0.88, 1.00)
$\rho_{c\Delta}$	0.71	0.01	(0.689, 0.735)	0.73	0.03	(0.684, 0.780)	0.69	0.01	(0.666, 0.712)
$\rho_{1\Delta}$	-0.29	0.04	(-0.37, -0.20)	-0.57	0.10	(-0.76, -0.38)	-0.22	0.08	(-0.37, -0.07)
$\rho_{2\Delta}$	-0.01	0.03	(-0.06, 0.04)	-0.01	0.14	(-0.27, 0.26)	-0.58	0.06	(-0.70, -0.46)
ρ_{1c}	0.19	0.05	(0.09, 0.29)	0.07	0.07	(-0.07, 0.22)	-0.07	0.04	(-0.15, 0.02)
ρ_{2c}	-0.61	0.02	(-0.65, -0.56)	-0.51	0.07	(-0.65, -0.38)	-0.14	0.04	(-0.21, -0.06)
ρ_{12}	-0.62	0.04	(-0.70, -0.54)	-0.65	0.07	(-0.78, -0.52)	0.08	0.04	(-0.00, 0.15)

standard deviations and correlation matrix for the random effects (see Table 2), state the full mathematical model used for the covariate adjusted analysis, and present dynamic individual prediction plots for the covariate adjusted model.

In Table 2 we give the estimated random effect standard deviations and correlations along with their standard errors and 95% percentile based confidence intervals. For the stratified analysis, there is increased variability in the first and second phase slopes (estimates of σ_{b1} and σ_{b2}) for lean women as compared to their obese counter-parts. The estimates of σ_Δ indicates that obese women show substantially more variability in when they arrive at the hospital relative to lean women. The variability in the change point was similar in both groups. The notable difference in the estimated correlation matrices is $\rho_{b_1 b_2}$, where the random effects for the first and second phase slopes are highly negatively correlated for lean women and essentially uncorrelated for obese women. Thus, there appears to be a regression to the mean phenomenon for lean women where those with slow (rapid) progression in the first phase, tend to have rapid (slow) progression in the second phase of labor. We do not see this phenomenon for the women in the obese group.

We now discuss the model formulation for the covariate adjusted analysis presented in Section 6.2 of the main text. For woman i, let Z_{i1} denote the indicator that woman i was in the obese pre-pregnancy BMI category, Z_{i2} the age of women i minus 24 (the median age), and $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})$. For simplicity we state the the full covariate adjusted model for $c_i > 0$, $c_i < 0$ follows from equation (1) in the main text. The covariate adjusted model for $Y_i(s)$ is

$$Y_i(s) = \begin{cases} K + (\beta_1 + \beta_{1\text{BMI}} Z_{i1} + \beta_{1\text{AGE}} Z_{i2})s + \epsilon_i(s) & \text{if } s < c_i \\ K + (\beta_1 + \beta_{1\text{BMI}} Z_{i1} + \beta_{1\text{AGE}} Z_{i2})c_i + (\beta_2 + \beta_{2\text{BMI}} Z_{i1} + \beta_{2\text{AGE}} Z_{i2})(s - c_i) + \epsilon_i(s) & \text{if } s \geq c_i \end{cases} \quad (5)$$

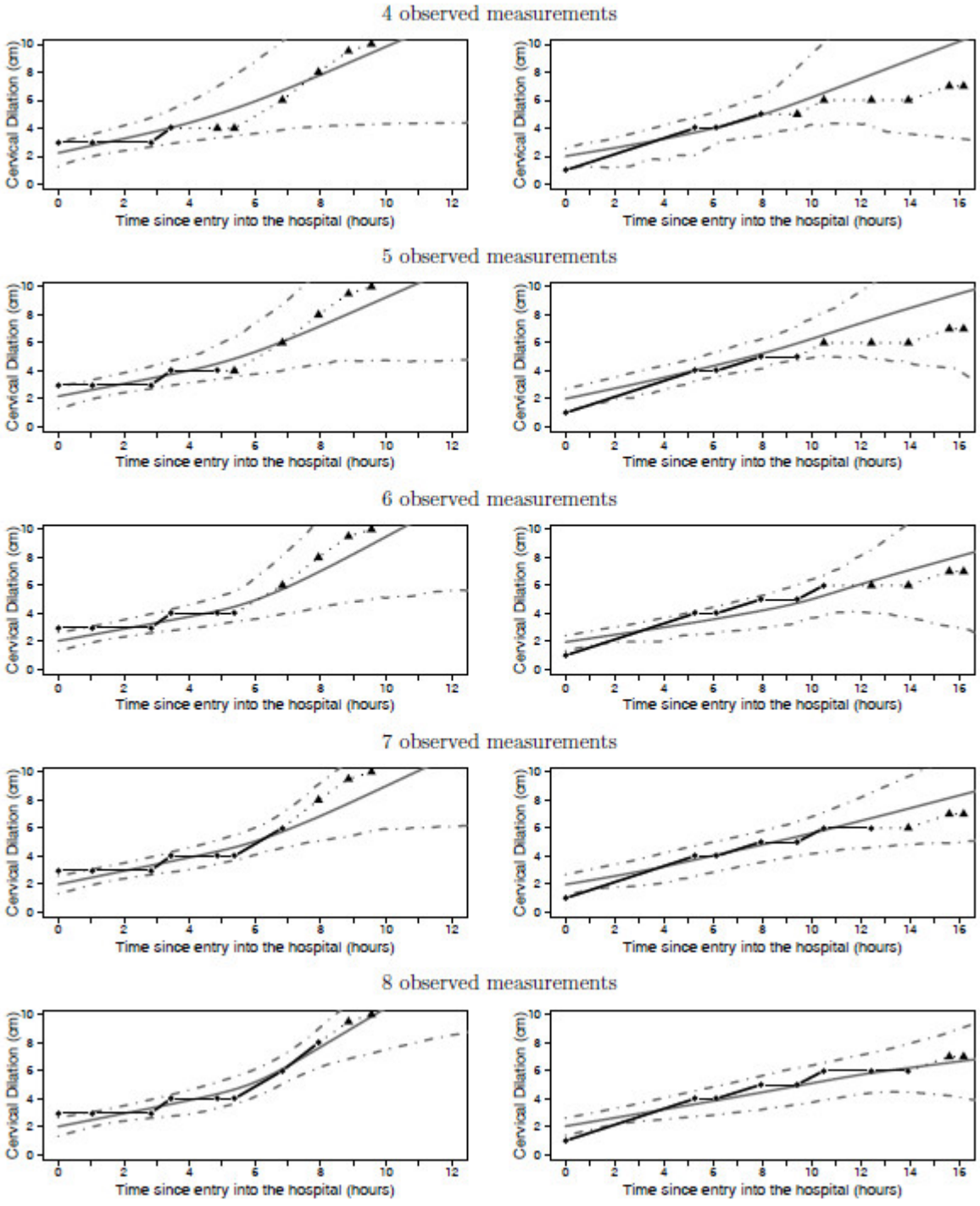
where

$$s = t - (\Delta + \Delta_{\text{BMI}} Z_{i1} + \Delta_{\text{AGE}} Z_{i2} + b_{i\Delta}),\ c_i = c + (c + c_{\text{BMI}} Z_{i1} + c_{\text{AGE}} Z_{i2} + b_{ic}),$$

$\boldsymbol{b}_i = (b_{i1}, b_{i2}, b_{i\Delta}, b_{ic}) \sim \text{MVN}(\mathbf{0}, \Sigma)$, and $_i(s) \sim^{iid} N(0, \sigma^2)$. It is straightforward to amend the analysis procedure to include the covariates.

In Figure 2 we display the predictions for a randomly chosen obese and lean woman by the number of observations in $\mathcal{D}_{ik}$. We display the estimated mean of the posterior distribution of the expected cervical dilation, along with the estimated 2.5th and 97.5th percentiles for both women. Using the notation in Section 4 of the main text, this corresponds to the estimates of $\tilde{\mu}(t|\mathcal{D}_{ik}, \hat{\boldsymbol{\theta}})$, $\tilde{Q}_{0.025}(t|\mathcal{D}_{ik}, \hat{\boldsymbol{\theta}})$, and $\tilde{Q}_{0.975}(t|\mathcal{D}_{ik}, \hat{\boldsymbol{\theta}})$ for $k = 4$, 5, 6, 7, and 8. Each figure shows the observed measurements $\mathcal{D}_{ik}$ as well as the unknown future observations. The estimate of $\tilde{\mu}$ gives the expected cervical dilation trajectory, while $Q_{0.025}$ and $Q_{0.975}$ give an estimate of the range for the expected trajectory for 95% of women with the given observed data $\mathcal{D}_{ik}$. The predictions were implemented using covariate adjusted model presented above. The left hand side of Figure 2 represents our random chosen lean woman, while the right side is the obese woman.

For the lean woman, her expected cervical dilation trajectory is relatively consistent by k. It is predicted that she be dilated a full 10 cm approximately 11 hours after she entered the hospital for all values of k (she reach 10 cm dilation at $t = 9.5$ hours). For the obese woman she did not reach the full 10 cm dilation. Her labor was stopped at 7 cm approximately 16 hours after she arrived at the hospital. With only the first 4 measurements, it is evident that this labor will longer than the lean woman's labor. It is clear, once observations 7 and 8 are included, that it is doubtful that this woman will reach full dilation before 16 hours.



Web Figure 2: Prediction of the expected cervical dilation (solid gray) with 95% prediction intervals (dot dash gray) by the size of the training sample for a randomly chosen lean (left) and obese (right) woman with 10 observations. The firgure constains the training data (solid black line with black dots) and the future observations (dashed line with black triangles).