



Published in final edited form as:

Stat Biosci. 2022 April ; 14(1): 1–22. doi:10.1007/s12561-021-09311-9.

Discriminatory capacity of prenatal ultrasound measures for large-for-gestational-age birth: A Bayesian approach to ROC analysis using placement values

Soutik Ghosal¹, Zhen Chen^{*,1}

¹Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, MD 20892

Abstract

Predicting large fetuses at birth is of great interest to obstetricians. Using an NICHD Scandinavian Study that collected longitudinal ultrasound examination data during pregnancy, we estimate diagnostic accuracy parameters of estimated fetal weight (EFW) at various times during pregnancy in predicting large-for-gestational-age. We adopt a placement value based Bayesian regression model with random effects to estimate ROC curves. The use of placement values allows us to model covariate effects directly on the ROC curves and the adoption of a Bayesian approach accommodates the *a priori* constraint that an ROC curve of EFW near delivery should dominate another further away. The proposed methodology is shown to perform better than some alternative approaches in simulations and its application to the Scandinavian Study data suggests that diagnostic accuracy of EFW can improve about 65% from week 17 to 37 of gestation.

Keywords

AUC; Estimated Fetal Weight; Obstetrics; Macrosomia; Diagnostic accuracy

1 Introduction

Excessive birth weight has been a concern for long in obstetrics (Esakoff et al., 2009), as newborns with large-for-gestational-age (LGA), defined as birth weight above the 90th percentile at a given gestation, can cause difficulty in delivery and lead to elevated likelihood of overweight in adulthood. As such, early prediction of LGA is of considerable clinical importance with potential to inform decisions on time and mode of delivery as well as on careful management of maternal dietary intake and weight gain. Repeated prenatal

Terms of use and reuse: academic research for non-commercial purposes, see here for full terms. <https://www.springer.com/aam-terms-v1>

* zhen.chen@nih.gov .

Publisher's Disclaimer: This Author Accepted Manuscript is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Data Accessibility

The data that support the findings of this study are available from the corresponding author, upon request.

ultrasound examinations have been widely used in this quest (Albert, 2012; Foster et al., 2017; Liu and Albert, 2014; Zhang et al., 2012).

The NICHD Successive Small-for-Gestational-Age Births Study in Scandinavia (Scandinavian Study hereafter) was designed to understand abnormal fetal growth and its risk factors, and provided rich data to undertake the aforementioned prediction task (Bakketeig et al., 1993). The Scandinavian Study enrolled multiparous women of Caucasian origin who spoke one of the Scandinavian languages and had a singleton pregnancy. All women were scheduled to have four ultrasound visits at 17, 25, 33 and 37 weeks of gestation. We are interested in the discriminatory capacity of Estimated Fetal Weight (EFW) from these ultrasound measurements in predicting LGA birth. In obstetrics, EFW is derived using ultrasound measurements of fetus, including head and abdominal circumferences and femur length (Hadlock et al., 1985). We wish to address the following questions using the Scandinavian Study data in this paper: What is the diagnostic accuracy of EFW at 17 weeks of gestation in relation to LGA birth? What are the corresponding diagnostic accuracy parameters at 25, 33, and 37 weeks? How about at 20, 30 weeks of gestation, or other similar times that are not among the scheduled? Is there any trend in these diagnostic accuracy estimates over the course of pregnancy?

Several challenges arise when we sought to answer these research questions using the ROC analysis techniques (Pepe, 2004). Although the four ultrasound examinations were scheduled to happen at weeks of 17, 25, 33, and 37, their actual times of occurrence only approximated the scheduled. In the remainder of the paper, we will use pGA and aGA to refer to the planned (scheduled) and actual times of ultrasound examinations respectively. As demonstrated by the density plots in Figure 1, there is substantial variability in aGAs. For example, the aGAs corresponding to 17-week of pGA has a range of 12 to 21.57 with a standard deviation of 1.25. This suggests that the approximation would not be reasonable if pGA were used in place of aGA. There are also considerable overlaps between neighboring densities in Figure 1, suggesting that ultrasound examinations scheduled for an earlier visit can happen after some of those scheduled for a later visit and vice versa. Furthermore, there are noticeable differences in the four distributions by pGA, with early ones more variable and some heavy tailed, indicative of a lack of uniform pattern in the differences between pGA and aGA over time. In the presence of these features, it is possible that the naïve analysis of grouping data according to pGA and applying a standard approach (such as a Bi-Normal model, (Metz, 1986)) to each of the four data subsets can produce estimates that are less accurate and less efficient.

It is well accepted in obstetrics that EFWs from ultrasound examinations closer to delivery have higher discriminating capacity for LGA than those from further (Ben-Haroush et al., 2007a,b; Bryant et al., 1997; Hedriana and Moore, 1994; Larsen et al., 1995; Pressman et al., 2000). Such *a priori* information, whenever available, should be incorporated in the statistical estimation process, since they can potentially produce more efficient estimates. Early work in literature have approached this task by considering *a priori* constraints in EFW distributions (e.g., (Chen and Hwang, 2019; Hwang and Chen, 2015)). While this practice is useful, it is more desirable and natural to apply constraints directly on ROC curves and

the associated area under the curve (AUC) measures. Yet, standard approaches such as the Bi-Normal model does not easily accommodate this need.

Motivated by these challenges in analyzing the Scandinavian Study data, we adopt a modeling framework for ROC curves and AUCs that builds on the placement value (PV), a term defined as the standardization of the diseased score with respect to the healthy score distribution. Within the context of our motivating example, the standard approaches to ROC analysis estimate ROC curves and AUCs by working with the EFW distributions of LGA and non-LGA populations separately (we refer to it as score-based approach thereafter). In contrast, the PV-based approach focuses on modeling the EFW distribution in the non-LGA population and the distribution of the PVs. As an ROC curve can be shown to be the cumulative distribution function (CDF) of the placement values, the PV-based framework provides a natural platform to model ROC curves and AUCs, especially when covariates are in consideration. First introduced by DeLong et al. (1988), PV-based ROC methodologies were furthered by Pepe (2000), Alonzo and Pepe (2002), Cai (2004), Pepe and Cai (2004), Stanley and Tubbs (2018), Inácio de Carvalho and Rodríguez-Álvarez (2018), among others. In the Scandinavian Study data, we will treat aGA as a covariate and construct a regression model of ROC curve as a function of it. In doing so, we are able to estimate diagnostic accuracy parameters not only for the scheduled weeks of gestation, but also for those not scheduled. Furthermore, given that it is directly modeling the ROC curves, this PV-based approach provides an easy and natural framework to incorporate the *a priori* constraint that ROC of EFWs later in gestation should dominate that of EFWs earlier.

This paper makes several contributions. First, it adopts a relative new but useful statistical framework to address the practical need of analyzing the Scandinavian Study data. In particular, the PV-based approach enables us to answer practical questions that would have been difficult using standard score-based approaches. The paper also considers constrained statistical analysis in ROC analysis and shows that the PV-based framework provides a straightforward way to accommodate ordering constraints on ROC curves, a difficult task with score-based framework. Recognizing the longitudinal data nature of the Scandinavian Study, this paper also considers random effects model in PV-based approach to explicitly accommodate dependent data; this will be illustrated in Section 3.3 and in the analysis of the Scandinavian Study data (Section 5). Finally, it addresses the issue of uncertainty in estimating PVs, a topic that has not received much attention in the literature on PV-based approaches to ROC curves.

The rest of the article is organized as follows. Section 2 presents a preliminary analysis of the Scandinavian Study data by a standard method, Section 3 provides detailed development of the PV-based approach to ROC analysis and discusses estimations of the model parameters. We demonstrate the performance of the developed methodology through simulation studies in Section 4 and present an application to the Scandinavian Study data in Section 5. We conclude with a brief discussion in Section 6.

2 An initial look at the Scandinavian Study data

The Scandinavian Study data consist of 2072 pregnant women, of which 190 do not have any ultrasound examination data, 6 have irregular visit time stamps, and 4 have missing birth outcomes. These 200 participants were removed from our analysis. In the remaining participants, 186 are classified to be LGA births, which are approximately the top 10% of the birth weights.

Since the ultrasound examinations were scheduled to happen at gestational age of 17, 25, 33, and 37 weeks, a natural starting point is to divide the data into four subsets according to these pGAs. Figure 2 presents boxplots of EFWs of such subsets stratified by LGA status. As pregnancy progresses, EFW increases in both LGA and non-LGA populations while the difference between them also gradually increases. To estimate diagnostic parameters of EFW in predicting LGA at these scheduled weeks, we apply a Bi-Normal model to each of the four subset data. The estimated ROC curves and AUCs are provided in Figure 3 and Table 1 respectively. We note that the AUC increases from 0.542 (95% CI 0.499 – 0.585) at week 17 to 0.841 (95% CI 0.806 – 0.871) at week 37.

These results are obtained based on the assumption that ultrasound examinations happened at the scheduled gestation age, which we learned from Figure 1 may not be reasonable. As it turns out, these estimates are less reliable because of the mis-alignment between the scheduled and actual times of examinations. In addition, such an analysis is not able to estimate diagnostic accuracy parameters at non-scheduled gestation weeks. In the next section, we show that a PV-based regression approach can avoid these pitfalls and can accommodate the *a priori* belief that EFW has higher discriminatory capacity when it is taken closer to delivery.

3 Modeling framework

3.1 Notations and background

A receiver operating characteristic (ROC) curve is a graphical tool to illustrate the diagnostic ability of a continuous test to discriminate two populations, say healthy and diseased (Pepe, 2004). In the context of the Scandinavian Study, let H and D be the status of the non-LGA and LGA birth, respectively and let y_i^H and y_i^D be EFWs of the corresponding fetus i ($i = 1, \dots, N_H$; N_H =number of non-LGA participants) and i ($i = 1, \dots, N_D$; N_D =number of LGA participants). Let y_i^H and y_i^D have cumulative distribution functions F_H and F_D , respectively. Then the ROC curve is defined as

$$ROC(t) = 1 - F_D(F_H^{-1}(1-t)), t \in (0, 1).$$

The area under the ROC curve is a measure interpretable as the probability that a randomly selected EFW of an LGA birth is higher than that of a non-LGA birth and is given as

$$AUC = \int_0^1 ROC(t)dt.$$

Different models of ROC arise with different specifications of F_D and F_H . The Bi-Normal model is a semiparametric approach with normality assumption in both diseased and healthy populations after an arbitrary monotone transformation τ ,

$$\tau(y_i^D) \sim N(\mu_D, \sigma_D^2) \text{ and } \tau(y_{i'}^H) \sim N(\mu_H, \sigma_H^2), i = 1, \dots, N_D, i' = 1, \dots, N_H.$$

The resultant ROC curve and AUC have closed form expressions (Metz, 1986)

$$ROC(t) = \Phi\left(a + b\Phi^{-1}(1-t)\right), t \in (0, 1),$$

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right),$$

where

$$a = \frac{\mu_D - \mu_H}{\sigma_D}, b = \frac{\sigma_H}{\sigma_D},$$

and $\Phi(\cdot)$ is the CDF of standard Normal distribution.

3.2 The PV-based approach to ROC

The use of placement values (PV) to make inference about ROC curve is not new (Hanley and Hajian-Tilaki, 1997). PV has been extensively used in several parametric, semi-parametric and non-parametric methods to estimate ROC curves (Alonzo and Pepe, 2002; Cai, 2004; Pepe and Cai, 2004; Qin and Zhou, 2006). A placement value is the standardization of a score from the diseased population relative to the distribution of the healthy scores. In particular, the PV of a diseased score Y_i^D is defined as

$$z_i = 1 - F_H(y_i^D).$$

Indeed, z_i can be interpreted as the proportion of healthy population with scores higher than y_i^D . It can be easily shown that the CDF of the random variable representing z_i is the ROC curve to discriminate healthy and diseased populations. As such, whereas we model y_i^D and $y_{i'}^H$ in the score-based approaches of ROC analysis, we model $Y_{i'}^H$ and z_i in the PV-based approach. This direct modeling approach accommodates direct covariate effects on ROC curves and allows global constraints to be incorporated if available. In the rest of the paper, we will term the two modeling components in PV-based approach, that of $y_{i'}^H$ and z_i as “Stage 1” and “Stage 2”, respectively.

To estimate PVs in Stage 1, we model EFWs of non-LGA births $y_{i'}^H \sim F_H$. Both parametric and non-parametric approaches can be used to model F_H . For example, we can specify

$$h(y_i^H) \sim N(\mu_H, \sigma_H^2), \quad (1)$$

where $h(\cdot)$ is an appropriate transformation function such as identity and logarithm function. Then the corresponding PV is estimated as

$$z_i = 1 - \hat{F}_H(y_i^D) = 1 - \Phi\left(\frac{h(y_i^D) - \hat{\mu}_H}{\hat{\sigma}_H}\right).$$

Since the CDF of the placement value z is equivalent to the ROC curve corresponding to the original scores, a variety of models can be fit to z , as detailed in the next subsection.

3.3 The PV-based Normal regression model

In the next few subsections, we introduce the PV-based normal regression model and sequentially incorporate the pieces necessary to answer the research questions we intend to answer in this manuscript.

3.3.1. Inclusion of covariates—The research questions in Scandinavian Study data outline in the introduction section necessitate a regression framework. ROC regression with covariates has been strongly advocated recently in the literature for valid diagnostic analysis and for associating discriminatory capacity measures to useful subject and test factors. Pepe (1997), Alonzo and Pepe (2002) and Dodd and Pepe (2003) represent some of the earliest parametric approaches related to placement values. Semi- and non-parametric extensions have been addressed in Cai and Pepe (2002), Cai (2004), Pepe and Cai (2004), Cai and Moskowitz (2004), and Lin et al. (2012), among others. Inácio de Carvalho and Rodríguez-Álvarez (2018) considered Bayesian semiparametric analysis of ROC regressions. As the objective of this paper is not to propose a new ROC regression framework, we content with the parametric approaches.

Let X_i be the vector of p covariates whose effects on ROC and AUC are of interest. For simplicity, we consider the case where covariates enter both the models for Y^H and Z while acknowledging that in general different covariates can be used. In the presence of covariates, the model in Stage 1 for Y^H can simply be extended to a multiple regression. For example, we can generalize model in equation (1) to

$$h(y_{i'}^H | X_{i'}) = X_{i'}^T \alpha + e_{i'}, e_{i'} \sim N(0, \sigma_H^2), i' = 1, \dots, N_H, \quad (2)$$

where α are the regression parameters. In this context, the PV specific to (y_i^D, X_i) is estimated as

$$z_i | X_i = 1 - \hat{F}_H(y_i^D | X_i) = 1 - \Phi\left(\frac{h(y_i^D) - X_i^T \hat{\alpha}}{\hat{\sigma}_H}\right).$$

In Stage 2, the model for z_i can involve a multiple regression as follows

$$\eta(z_i | X_i) = X_i^T \beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), i = 1, \dots, N_D, \quad (3)$$

where $\eta(x)$ is a link function that maps the PVs from unit interval to the real line, β are regression coefficients, and σ^2 is the variance parameter for PV on η transformation scale. Once the regression coefficients are estimated ($\hat{\beta}$ and $\hat{\sigma}$), for a given value of covariate x_0 we can estimate ROC at x_0 as

$$\widehat{ROC}_{x_0}(t) = \Phi\left(\frac{\eta(t) - \hat{\mu}_{x_0}}{\hat{\sigma}}\right), t \in (0, 1), \text{ where } \hat{\mu}_{x_0} = x_0^T \hat{\beta},$$

i.e. the CDF of $N(\hat{\mu}_{x_0}, \hat{\sigma}^2)$ computed at points $\eta(t)$, $t \in (0, 1)$.

Several forms of η can be used, including the familiar logit, probit and logarithmic functions, among others. Chen and Ghosal (2020) compared the performance of PV-based approach to ROCs using these link functions under an array of mechanisms of generating the test scores (e.g., Bi-Normal, Bi-Mixture-Normal, Bi-Gamma), and found that both probit and logit link functions perform very well and are superior to other alternatives. A Beta model (Stanley and Tubbs, 2018) was recently proposed but is also shown to be inferior to logit/probit transformation models (Chen and Ghosal, 2020). Furthermore, the Beta regression model has poor Markov chain Monte Carlo convergence and mixing in our experience. For these reasons, we choose $\eta(x) = \Phi(x)$ in this paper.

3.3.2 Handling longitudinal data—When repeated measures $\{y_{i'j}^H, y_{i'j}^D\}$ are available, $i' = 1, \dots, N_H$; $i = 1, \dots, N_D$; $j = 1, \dots, K$, where K is the number of ultrasound examinations, a random intercept ($v_{i'}$) can be introduced in equation (2) to account for the potential correlations in estimating PV in Stage 1

$$h(y_{i'j}^H) = v_{i'} + W_{i'j}^T \alpha + e_{i'j}, e_{i'j} \sim N(0, \sigma_H^2), i' = 1, \dots, N_H, \quad (4)$$

where $v_{i'} \sim N(0, \sigma_v^2)$. The corresponding PV estimator is given by

$$z_{ij} | X_{ij} = 1 - \Phi\left(\frac{h(y_{ij}^D) - X_{ij}^T \hat{\alpha}}{\sqrt{\hat{\sigma}_H^2 + \hat{\sigma}_v^2}}\right).$$

Similarly we can introduce a random intercept in the equation (3) to account for correlation in modeling z_j in Stage 2:

$$\eta(z_{ij} | X_{ij}) = u_i + X_{ij} \beta + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2), i = 1, \dots, N_D; j = 1, \dots, K, \quad (5)$$

where $u_i \sim N(0, \sigma_u^2)$ is the random intercept.

3.3.3 Incorporation of constraints—Because it connects ROC curves to CDFs of placement values, the PV-based approach provides a natural framework to incorporate a

priori constraints. For example, in the Scandinavian Study data, it is believed that EFWs have higher discriminatory capacity in later ultrasound examinations. This prior belief can be translated into increasing AUCs or stochastically ordered placement values over time. This *a priori* constraint can be accommodated by specifying a truncated prior for components of β in equation (5). For example, in the case of a single covariate (aGA), (5) becomes

$$\eta(z_{ij} | X_{ij}) = u_i + \beta_0 + \beta_1 aGA_{ij} + \epsilon_{ij}. \quad (6)$$

Then we can constrain β_1 to be negative to reflect the prior belief, i.e.,

$$\beta_1 \sim N(\beta_{10}, \sigma_{\beta_1}^2) I(\beta_1 < 0),$$

where $I(c)$ is the usual indicator function that takes the value 1 if c is true and 0 otherwise.

3.4 Estimation, inference, and computation

We take a Bayesian approach to inference, which is particularly useful when *a priori* constraints are to be incorporated in parameter estimations. For priors, we use proper yet vague distributions to allow data to dominate the estimation. In particular, each of the regression coefficient vectors α , β follow $N(0, 100)$ priors and inverse of variance parameters σ^2 , σ_u^2 and σ_v^2 follow $Gamma(0.01, 0.01)$.

We use RJAGS to implement Markov Chain Monte Carlo (MCMC) algorithms to generate samples from the posterior distribution of the model parameters given the data. Both visual inspection of the trace plots and diagnostic tools (Gelman et al., 1992) are used to ensure convergence of the MCMC chains. After convergence, we thin the iterations to produce a sample of 5000 to produce posterior means, standard deviations and 95% credible intervals. R code of implementing simulation and real data analysis will be made available online.

Since the PV-based approach estimates PVs in Stage 1 before models them in Stage 2, it is necessary to appropriately account for the uncertainty associated with the first step. We achieve this by using bootstrap as follows:

- i. Generate a sample of EFWs, with replacement, from the non-LGA group. This step is implemented separately at each pGA;
- ii. Fit Stage 1 model to the EFW sample from (i). Then obtain PV estimates using posterior means of model parameters in equation (4);
- iii. Fit Stage 2 model to the PVs from (ii). Obtain posterior mean of the AUCs using L iterations of the MCMC algorithm;
- iv. Repeat steps (i)-(iii) R times. The average posterior means of AUCs (over R replications) is used as the final AUC estimates and the 2.5 and 97.5 percentiles used as the final variability estimates.

The idea is to construct an ensemble of R ROC curves (AUC values), each of which corresponding to the posterior mean obtained from one non-LGA EFW sample that is drawn

(with replacement) from the data. We note that this algorithm is essentially the same as the Bayesian bootstrap (Rubin, 1981), which has been used in ROC analysis to obtain smooth ROC curve and associated credible bands (Gu et al., 2008; Inácio de Carvalho and Rodríguez-Álvarez, 2018) when nonparametric approach was taken.

4 Simulation studies

4.1 Consequence of using pGA in place of aGA

We conduct simulation studies to assess the performance of the PV-based modeling framework when the actual ultrasound examination times (aGA) are used as a continuous covariate (referred to as “aGA-con” thereafter). We use as comparison some “naïve” alternatives that treat scheduled (planned) ultrasound examination times (pGA) as actual times. Three such alternatives are considered, one using pGA as continuous (“pGA-con”), second using it as categorical (“pGA-cat”), and the third using the naïve Bi-Normal analysis. The second “naïve” alternative relaxes the linearity assumption in the first. In the primary simulation scenario, we mimic the Scandinavian Study data by using similar aGAs at each pGA and by keeping the same sample size. In particular, we generate EFW data according to the following system

$$Y_{i'j}^H = a_1 + b_1 X_{i'j}^H + \epsilon_{i'j}^H, i' = 1, \dots, N_H,$$

$$Y_{ij}^D = a_2 + b_2 X_{ij}^D + \epsilon_{ij}^D, i = 1, \dots, N_D, \quad (7)$$

where $j = 1, \dots, 4$, $X_{i'j}^H$ and X_{ij}^D are times of ultrasound examinations sampled from normal distributions with means and variances matching those from the empirical distributions of aGA at each pGA (“Equal” variability) from the Scandinavian Study data and $\epsilon_{i'j}^H \stackrel{\text{ind}}{\sim} N(0, \sigma_H^2)$, $\epsilon_{ij}^D \stackrel{\text{ind}}{\sim} N(0, \sigma_D^2)$. We consider the scenarios where the true AUCs are 0.641, 0.702, 0.758 and 0.784 for examination times of 17, 25, 33 and 37, respectively. For brevity, we only consider the case of no-constraint and disregard the uncertainty in PVs, but note that the corresponding results are similar. We create 1000 data replicates and estimate the PVs for each dataset. Then we apply the four fitting models, aGA-con, pGA-con, pGA-cat, and naïve Bi-Normal model, to each PV dataset and estimate AUC at 17, 25, 33, and 37 weeks respectively. We calculate average bias (aBias), average standard deviation (aSD), and 95% credible interval based on the 1000 replications.

To assess the impact of the degree of misalignment between aGA and pGA on ROC and AUC estimates, we considered two additional situations where new aGAs are generated at each pGA from a normal distribution with same mean and variance double (“Double” variability) or half (“Half” variability) of the empirical counterparts in the Scandinavian Study data. The above primary situation is termed as “Equal” sample size. Finally, we also considered two more sample size situations, one large (two times of the original sample size at each pGA) and the other small (fifty percent of the original sample size at each pGA).

Simulation results from these 9 scenarios are reported in Tables 2–4 for the three sample size scenarios (each with three variability scenarios), respectively.

Focusing on the primary simulation scenario under the original sample size (middle panel of Tables 2), we see that AUC estimates from the adopted approach (aGA-con) are essentially unbiased (bias range: -0.002 , -0.003). The average posterior standard deviation is slightly higher at week 17 than at other time points, likely due to the slightly larger variability of aGA around pGA of 17. In comparison, the alternative approach that uses pGA as a continuous covariate (pGA-con) shows biases in AUC estimates from 0.03 (at week 17) to 0.17 (at week 37). These represent relative biases of 5.3% to 22.2%, respectively. The variability measures are comparable with those from aGA-con. Using categorical version of pGA (pGA-cat) produces very similar relative biases. The slightly enhanced variability estimates in pGA-cat can be a result of loss of efficiency due to grouping EFWs according to pGA. The results from the naïve Bi-Normal model are similar to those from pGA-cat, indicating that the estimates from the naïve Bi-Normal are biased and less efficient compared to the adopted approach (aGA-con).

Doubling and halving the variabilities of aGAs around pGAs impacted AUC estimates, with the former increase and the latter decrease biases in AUC estimates. For example, the bias of AUC estimate at week 37 is -0.174 (22.2%) under “Equal” variability scenario but increases to -0.226 (28.8%) under “Double” scenario. On the other hand, the corresponding bias decreases to -0.098 (12.5%) under “Half” scenario. These are both expected, as the variability of aGA around pGA represents the degree of misalignment between the two versions of examination times. The posterior standard deviations of AUC estimates appear to remain the same under either doubling or halving variability scenario.

Tables 3 and 4 depict similar pictures on the biases under the large and small sample size scenarios. As expected, the average standard deviations become smaller when the sample size is two times of the original and larger when sample size is half of the original.

Put together, these simulation results suggest that using pGA in place of aGA can produce biased AUC estimates. Hence, a naïve ROC analysis is not appropriate. The biases increase with the variability of aGA at pGA. These conclusions hold whether pGA is used as continuous or categorical or a stratified naïve Bi-Normal model is used. The conclusions also hold for all sample sizes.

4.2 Consequence of accommodating *a priori* constraints

In this simulation study we intend to study the efficiency gain due to inclusion of constraints. We generate covariate X from a $N(27.64, 8^2)$ distribution for 150 subjects. Then we generate the logit-transformed PVs as

$$Z^* = -0.5 - 0.01 \cdot X + \epsilon$$

where $\epsilon \sim N(0, 1)$, and obtain PVs

$$Z = \text{expit}(Z^*),$$

where *expit* is the familiar inverse logit function. We fit the adopted Normal regression models in section 3.3.3, equation (6) with and without constraint $\beta_1 < 0$ and calculate AUCs at $X = 37$. We repeat the process 1000 times and report average posterior means and standard deviations in Table 5. We can see that inclusion of constraints results in efficiency gains as the average posterior standard deviation changes from 0.0236 to 0.0201 when a *priori* constraint is embedded in the estimation.

5 A revisit of the Scandinavian Study data

5.1 Stage 1: estimating PVs

In order to estimate PVs, a first step is to model EFW distribution in the non-LGA population. To account for non-linearity and heteroskedasticity, we transform the response $Y^H = \sqrt[4]{\text{EFW}}$ and model $\log(Y^H)$ as a polynomial function of aGA with the error term as a function of aGA as well. We consider several mean and variance specifications and use the Deviance Information Criterion (DIC) to select the best; see Table 6 Stage 1 block. This leads us to a model

$$\log(y_i^H) = \alpha_0 + \alpha_1 aGA_i + \alpha_2 aGA_i^2 + \epsilon_i, \epsilon_i \sim N(0, \sigma_H^2). \quad (8)$$

PV z_i are then calculated as

$$z_i = 1 - \Phi\left(\frac{\log(y_i^D) - \hat{\alpha}_0 - \hat{\alpha}_1 aGA_i - \hat{\alpha}_2 aGA_i^2}{\hat{\sigma}_H}\right),$$

where the same quad-root transformation is applied to y_i^D . Residual plots of this modeling process is provided in Figure 4 which demonstrates satisfactory fit.

5.2 Stage 2: estimating ROC curves & AUCs

We repeat a similar model selection process as above to choose an appropriate model for the placement values $Z = (z_1, \dots, z_{ND})$. Based on the scatter plot of Z in Figure 5 and model selection criteria in Table 6 Stage 2 block, we fit the following model to Z

$$\Phi(z_i) = \beta_0 + \beta_1 aGA_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2).$$

Residual plots of this stage of modeling also show satisfactory model fitting (Figure 5).

Using the fitted model, we estimate ROC curves and AUCs at scheduled gestation ages as well as two unscheduled weeks (20 and 30). The AUC estimates are presented in Table 7 (model A). However, these analyses assumed uncorrelated EFWs despite the repeated measures in the dataset. Indeed, a cursory examination of correlations between EFWs at different pGAs reveals that the correlations range from 0.11 to 0.64. To adequately account for these correlations, we introduce random intercepts in the above models similar to equation (6). The estimated AUCs are reported in Table 7 (model B). We notice that

while AUC estimates in this model are very similar to those of model A, posterior standard deviations are relatively larger, which is expected.

The results when we further impose the *a priori* constraint that later EFWs have higher AUCs are presented in model C in Table 7. The results are almost the same as those from model B, likely due to the fact that the Scandinavian Study data already agree with the *a priori* constraint. Indeed, the estimated ROC curves in Figure 6 clearly demonstrate a strict ordering in the sense that ROCs at later pregnancy (e.g., 33 weeks) dominate those at early pregnancy (e.g., 30 weeks).

To account for uncertainty in estimating the PVs we fit the model using bootstrap method described in Section 3.4 with $R = 1000$ resampling replicates. AUC and ROC curves estimates are presented in Table 7 (model D) and Figure 6, respectively. The AUC estimates are essentially the same as in model C. However, the posterior SDs have increase slightly, as a result of adequately reflecting the uncertainty in Stage 1 modeling which models A through C have so far ignored.

As comparisons, we also fit two alternatives that mimic the “pGA-con” and “pGA-cat” models in Section 4.1. In particular, the first alternative (model E thereafter) is the same as model D, except that aGA is replaced with the continuous pGA in both Stages 1 and 2. In the second alternative (model F thereafter), we use categorical pGA in place of aGA in model D and also let the error term depend on pGA. Since pGA is categorical in this model, the quadratic specification in Stage 1 is not necessary, and AUCs at un-scheduled week of gestation (e.g., 20 and 30) can not be estimated. The heteroschedasticity specification of model F also makes it difficult to consider the *a priori* constraint. So we disregarded it.

The AUC estimation results of these two alternative models are reported in the bottom two rows of Table 7. Compared to results from models E and F, model D estimates higher AUCs at the six times considered except at 17 weeks. This suggests that the naïve alternative models tend to underestimates AUCs, especially at later gestation weeks. The below-chance AUC estimate at week 17 under model E could be a result of the stringent linearity assumption. The relative smaller posterior SD estimates in model E are likely the result of using pGA which has considerably smaller variability than aGA. On the other hand, the slightly larger posterior SD estimates in model F is an indication of loss of statistical efficiency due to effectively divide the original data into smaller subsets.

6 Discussion

The use of EFWs during pregnancy to predict LGA at birth is clinically important as it informs decisions on delivery routes as well as maternal dietary and weight management. Our analysis of the Scandinavian Study data suggests that the diagnostic accuracy as measured by the AUC can increase from slightly above 0.5 at 17 week to around 0.86 at 37 week of gestation.

Adopting a regression framework based on placement values, we were able to obtain these diagnostic accuracy estimates at both scheduled and unscheduled visits. We showed that noticeable changes in estimates can happen when we account for correlations in the

data, incorporate *a priori* constraints, and accommodate uncertainty in estimating placement values.

We have focused on parametric specifications in both PV and AUC estimations in this paper. Although satisfactory in the analysis of the Scandinavian Study data, these parametric approaches can be restrictive. Future work are warranted to develop semi-parametric and non-parametric approaches. For example, we can consider Dirichlet process mixtures in modeling Y^H and Z .

As useful as it is in discriminating “healthy” and “diseased” populations using a biomarker, the machinery of ROC curve analysis might not be the most relevant in some situations, especially when prediction is the primary goal. It is of interest to investigate how predictiveness curve (Gail and Pfeiffer, 2005) can be utilized in the context of EFW and LGA and how the analytical challenges illustrated in the paper can be handled.

Some other future work include extensions of the current approach to situations where the outcome is ordinal, e.g. small for gestation age (SGA), normal, and LGA, and consideration of proper ROC curves that guarantee concave ROC curves.

Acknowledgements

This research was supported by the Intramural Research Program of *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. This work utilized the computational resources of the NIH HPC Biowulf cluster. (<http://hpc.nih.gov>)

References

- Albert PS (2012). A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statistics in Medicine*, 31(2):143–154. [PubMed: 22081439]
- Alonzo TA and Pepe MS (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3(3):421–432. [PubMed: 12933607]
- Bakketeig LS, Jacobsen G, Hoffman HJ, Lindmark G, Bergsjø P, Molne K, and Rødsten J (1993). Pre-pregnancy risk factors of small-for-gestational age births among parous women in Scandinavia. *Acta Obstetrica et Gynecologica Scandinavica*, 72(4):273–279. [PubMed: 8389514]
- Ben-Haroush A, Chen R, Hadar E, Hod M, and Yogev Y (2007a). Accuracy of a single fetal weight estimation at 29–34 weeks in diabetic pregnancies: Can it predict large-for-gestational-age infants at term? *American Journal of Obstetrics and Gynecology*, 197(5):497.e1–497.e6. [PubMed: 17980186]
- Ben-Haroush A, Yogev Y, Hod M, and Bar J (2007b). Predictive value of a single early fetal weight estimate in normal pregnancies. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 130(2):187–192. [PubMed: 16720074]
- Bryant DR, Zador I, Landwehr JB, and Wolfe HM (1997). Limited clinical utility of midtrimester fetal morphometric percentile rankings in screening for birth weight abnormalities. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 177(4):859–863.
- Cai T (2004). Semi-parametric ROC regression analysis with placement values. *Biostatistics*, 5(1):45–60. [PubMed: 14744827]
- Cai T and Moskowitz CS (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics*, 5(4):573–586. [PubMed: 15475420]
- Cai T and Pepe MS (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97(460):1099–1107.

- Chen Z and Ghosal S (2020). A note on modeling placement values in the analysis of receiver operating characteristic curves. *Biostatistics & Epidemiology*, pages 1–16, DOI: 10.1080/24709360.2020.1737794.
- Chen Z and Hwang BS (2019). A Bayesian semiparametric approach to correlated ROC surfaces with stochastic order constraints. *Biometrics*, 75(2):539–550. [PubMed: 30390405]
- DeLong ER, DeLong DM, and Clarke-Pearson DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845. [PubMed: 3203132]
- Dodd LE and Pepe MS (2003). Partial AUC estimation and regression. *Biometrics*, 59(3):614–623. [PubMed: 14601762]
- Esakoff TF, Cheng YW, Sparks TN, and Caughey AB (2009). The association between birthweight 4000 g or greater and perinatal outcomes in patients with and without gestational diabetes mellitus. *American Journal of Obstetrics and Gynecology*, 200(6):672.e1–672.e4. [PubMed: 19376489]
- Foster JC, Liu D, Albert PS, and Liu A (2017). Identifying subgroups of enhanced predictive accuracy from longitudinal biomarker data by using tree-based approaches: Applications to fetal growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):247–261.
- Gail MH and Pfeiffer RM (2005). On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239. [PubMed: 15772102]
- Gelman A, Rubin DB, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.
- Gu J, Ghosal S, and Roy A (2008). Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine*, 27(26):5407–5420. [PubMed: 18613217]
- Hadlock FP, Harrist R, Sharman RS, Deter RL, and Park SK (1985). Estimation of fetal weight with the use of head, body, and femur measurements—a prospective study. *American Journal of Obstetrics and Gynecology*, 151(3):333–337. [PubMed: 3881966]
- Hanley JA and Hajian-Tilaki KO (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Academic Radiology*, 4(1):49–58. [PubMed: 9040870]
- Hedriana HL and Moore TR (1994). A comparison of single versus multiple growth ultrasonographic examinations in predicting birth weight. *American Journal of Obstetrics and Gynecology*, 170(5):1600–1606. [PubMed: 8203416]
- Hwang BS and Chen Z (2015). An integrated Bayesian nonparametric approach for stochastic and variability orders in ROC curve estimation: An application to endometriosis diagnosis. *Journal of the American Statistical Association*, 110(511):923–934. [PubMed: 26839441]
- Inácio de Carvalho V and Rodríguez-Álvarez MX (2018). Bayesian nonparametric inference for the covariate-adjusted ROC curve. *arXiv preprint arXiv:1806.00473*.
- Larsen T, Greisen G, and Petersen S (1995). Prediction of birth weight by ultrasound-estimated fetal weight: a comparison between single and repeated estimates. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 60(1):37–40. [PubMed: 7635228]
- Lin H, Zhou X-H, and Li G (2012). A direct semiparametric receiver operating characteristic curve regression with unknown link and baseline functions. *Statistica Sinica*, 22(4):1427–1456. [PubMed: 30799913]
- Liu D and Albert PS (2014). Combination of longitudinal biomarkers in predicting binary events. *Biostatistics*, 15(4):706–718. [PubMed: 24831103]
- Metz CE (1986). ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–733. [PubMed: 3095258]
- Pepe MS (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84(3):595–608.
- Pepe MS (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56(2):352–359. [PubMed: 10877289]
- Pepe MS (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.
- Pepe MS and Cai T (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60(2):528–535. [PubMed: 15180681]

- Pressman EK, Bienstock JL, Blakemore KJ, Martin SA, and Callan NA (2000). Prediction of birth weight by ultrasound in the third trimester. *Obstetrics & Gynecology*, 95(4):502–506. [PubMed: 10725480]
- Qin G and Zhou X-H (2006). Empirical likelihood inference for the area under the ROC curve. *Biometrics*, 62(2):613–622. [PubMed: 16918927]
- Rubin DB (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.
- Stanley S and Tubbs J (2018). Beta regression for modeling a covariate adjusted ROC. *Science Journal of Applied Mathematics and Statistics*, 6(4):110–118.
- Zhang J, Kim S, Grewal J, and Albert PS (2012). Predicting large fetuses at birth: Do multiple ultrasound examinations and longitudinal statistical modelling improve prediction? *Paediatric and Perinatal Epidemiology*, 26(3):199–207. [PubMed: 22471679]

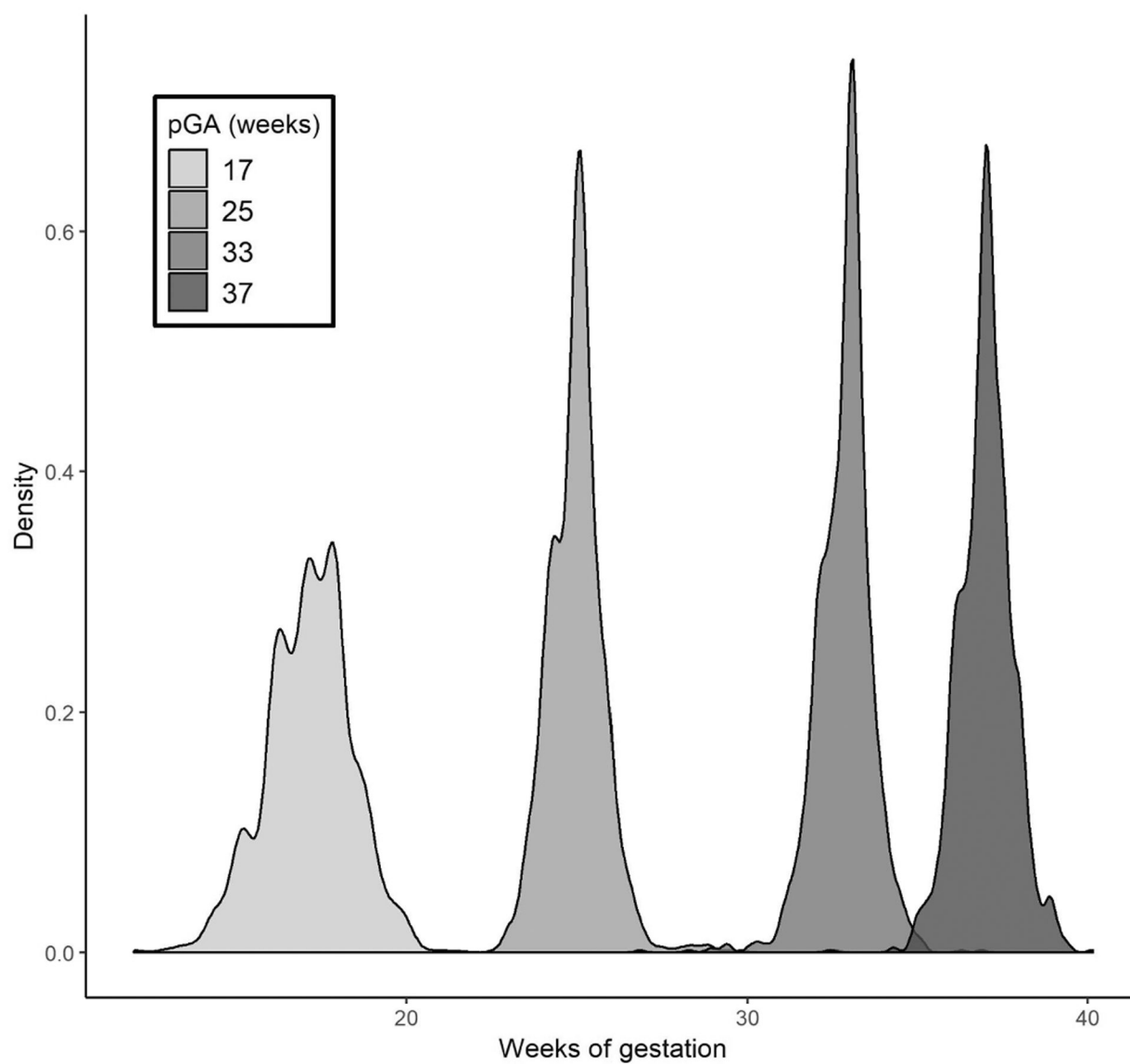


Figure 1:
Density plots of aGA, the gestation age (in weeks) when ultrasound examinations were actually taken, stratified by pGA, the gestation age (in weeks) when ultrasound examinations were scheduled to happen, the Scandinavian Study.

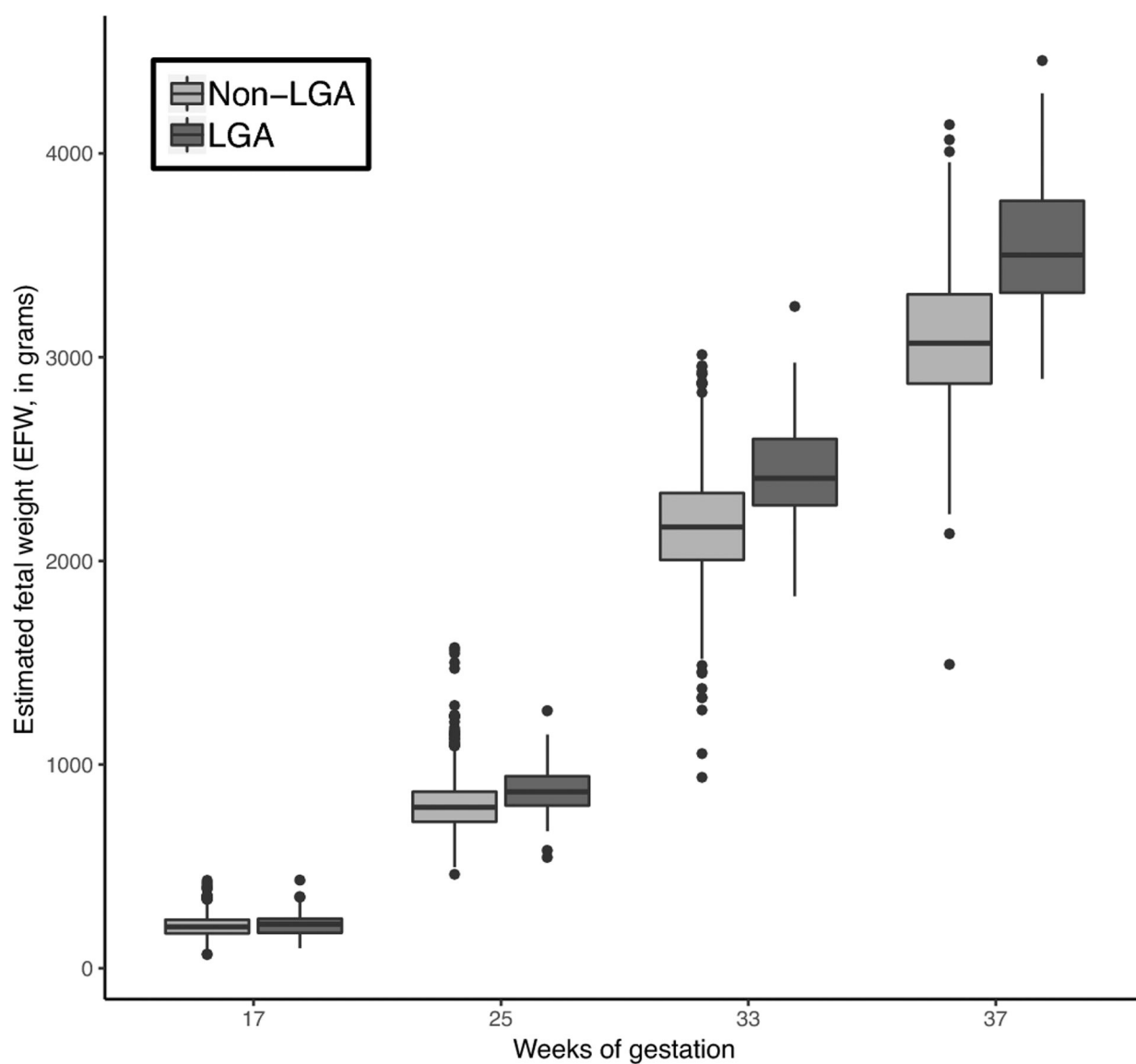


Figure 2:
Boxplots of EFW stratified by LGA status and pGA, the gestation age (in weeks) when ultrasound examinations were scheduled to take, the Scandinavian Study.

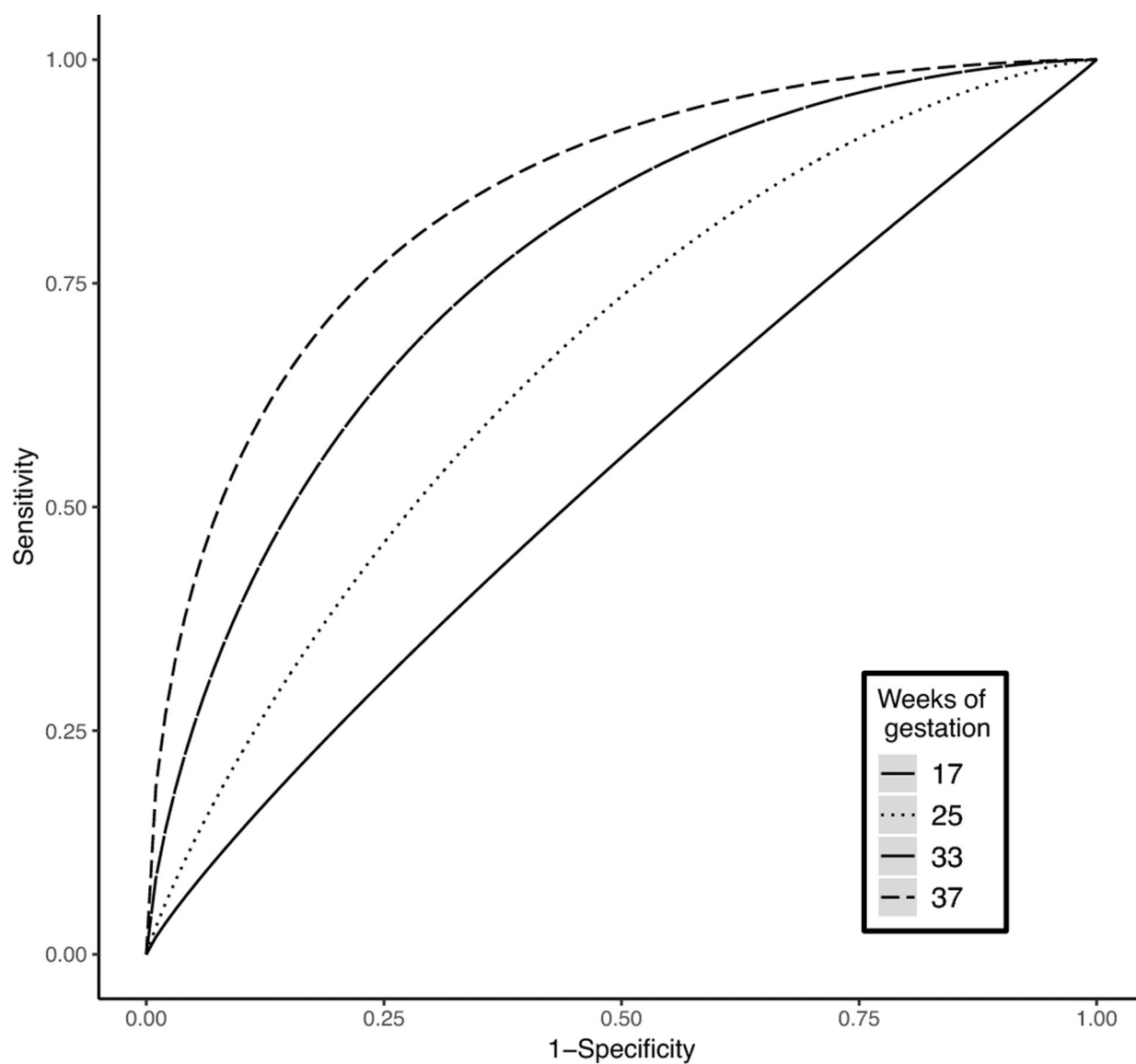


Figure 3:
Estimated ROC curves from a naïve Bi-Normal model, the Scandinavian Study.

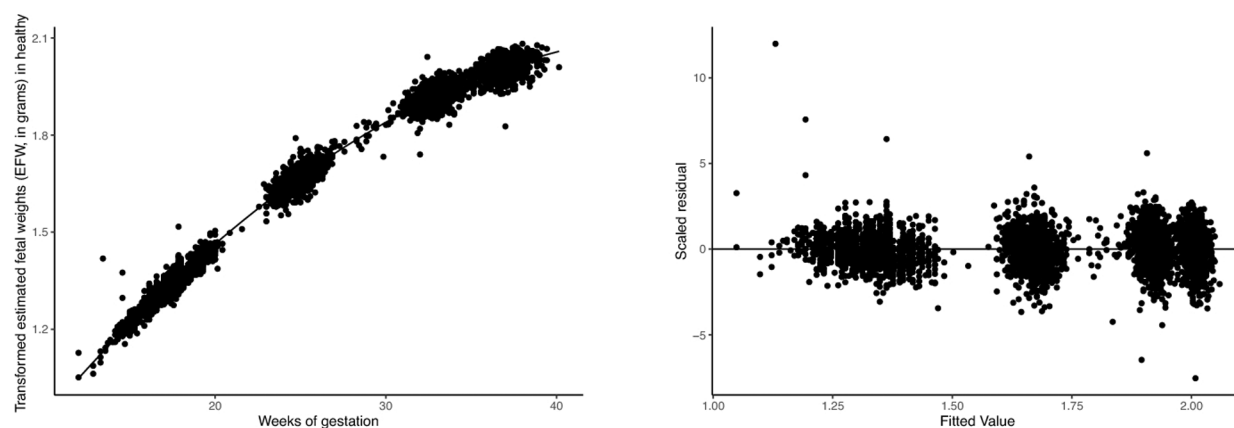


Figure 4: Model diagnostic plots of Stage 1 modeling process, the Scandinavian Study. The plots are based on the best model in Table 6, Stage 1 block. On the left panel, the Y-axis corresponds to logarithm of the non-LGA EFW after the quad-root transformation.

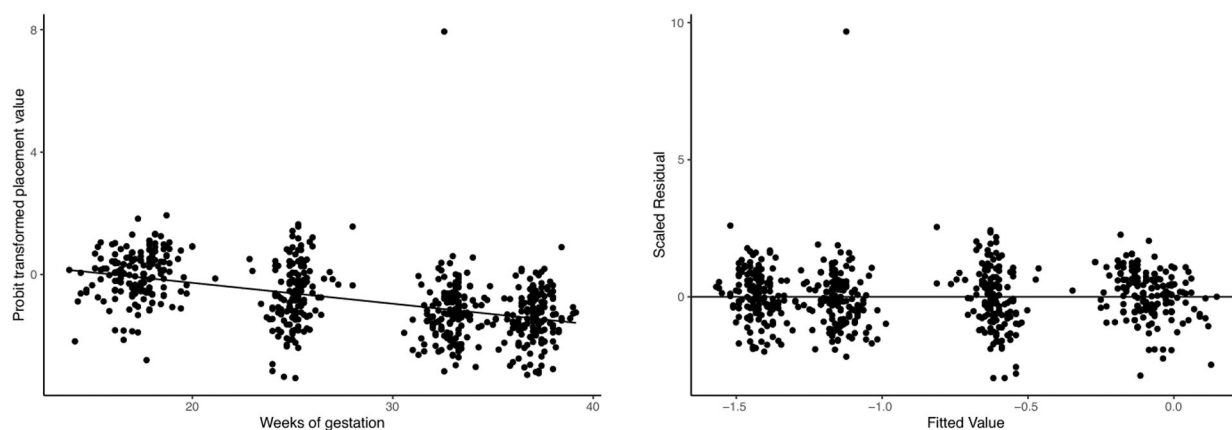


Figure 5:
Model diagnostic plots of Stage 2 modeling process, the Scandinavian Study. The plots are based on the best model in Table 6, Stage 2 block.

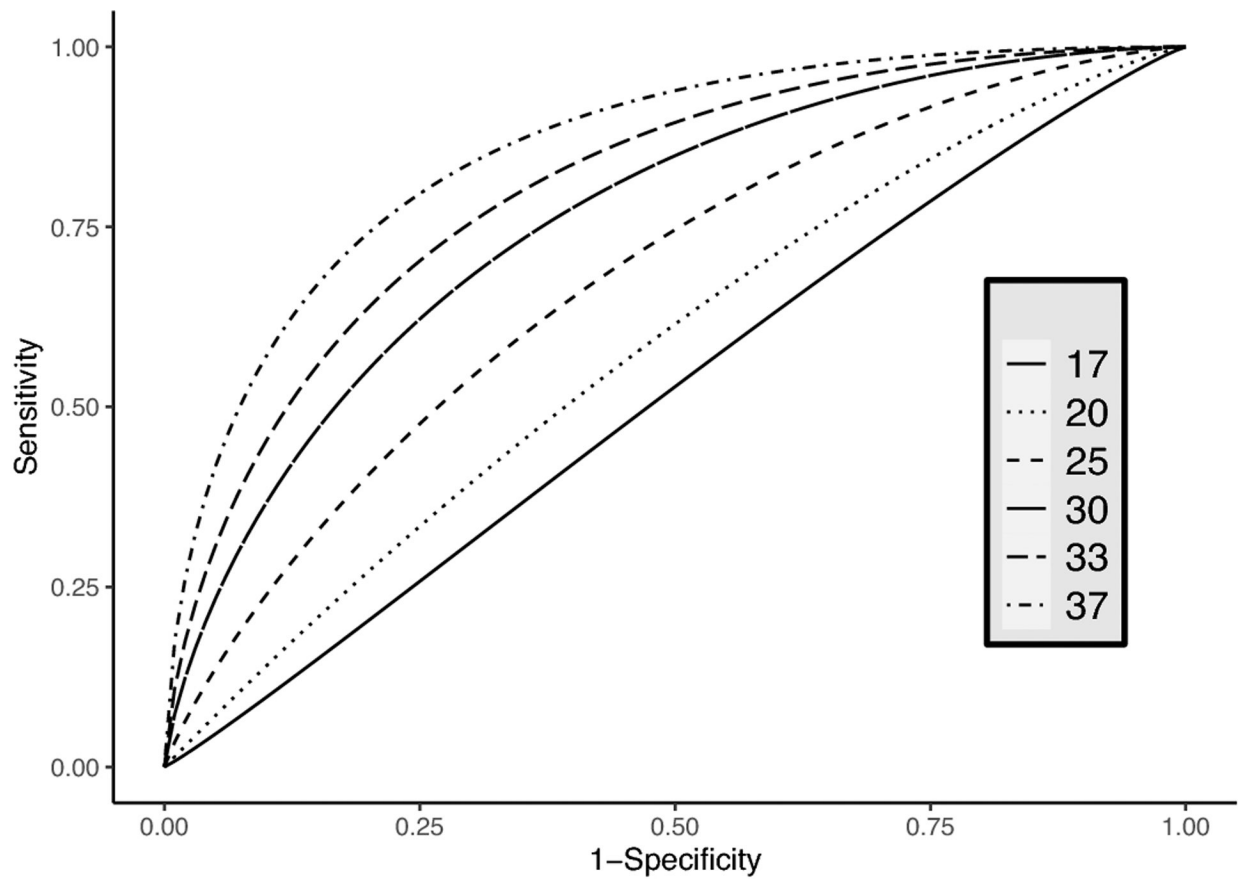


Figure 6:

Estimated ROC curves from the PV-based regression model with aGA as the continuous covariate (model D). It uses random intercepts to account for correlated data, imposes the *a priori* constraint that EFW at later stage of pregnancy has higher discriminatory capacity, and accounts for variability in estimate placement values in Stage 1, the Scandinavian Study data.

Table 1:

Posterior estimates of areas under ROC curves from the naïve Bi-Normal model, the Scandinavian Study.

	Weeks of gestation			
	17	25	33	37
Mean	0.542	0.664	0.773	0.841
SD \times 10	0.222	0.219	0.189	0.168
95% CI	(0.499, 0.585)	(0.620, 0.706)	(0.734, 0.808)	(0.806, 0.871)

Means and precisions of the normal distributions follow $N(0, 100)$ and $Gamma(0.01, 0, 01)$ priors, respectively.

Table 2:

Simulation results when sample size is same as that in the Scandinavian Study data.

Variability	Weeks	True AUC	aGA-con			pGA-con			pGA-cat			naïve Bi-Normal		
			aBias	aSD	95% CI	aBias	aSD	95% CI	aBias	aSD	95% CI	aBias	aSD	95% CI
Double	17	0.641	-0.001	0.010	(0.608, 0.675)	-0.084	0.011	(0.511, 0.598)	-0.087	0.013	(0.497, 0.608)	-0.101	0.013	(0.497, 0.582)
	25	0.702	-0.002	0.006	(0.681, 0.721)	-0.145	0.007	(0.532, 0.584)	-0.139	0.013	(0.523, 0.603)	-0.136	0.013	(0.524, 0.608)
	33	0.758	-0.003	0.006	(0.735, 0.775)	-0.200	0.008	(0.537, 0.579)	-0.204	0.013	(0.519, 0.588)	-0.194	0.012	(0.523, 0.607)
	37	0.784	-0.003	0.007	(0.756, 0.804)	-0.226	0.010	(0.530, 0.585)	-0.226	0.013	(0.522, 0.593)	-0.215	0.013	(0.526, 0.610)
Equal	17	0.641	-0.002	0.010	(0.605, 0.675)	-0.034	0.011	(0.563, 0.649)	-0.040	0.012	(0.546, 0.653)	-0.062	0.012	(0.535, 0.620)
	25	0.702	-0.002	0.006	(0.681, 0.721)	-0.094	0.007	(0.582, 0.634)	-0.082	0.012	(0.582, 0.659)	-0.077	0.013	(0.585, 0.665)
	33	0.758	-0.003	0.006	(0.735, 0.776)	-0.149	0.007	(0.588, 0.631)	-0.155	0.012	(0.569, 0.639)	-0.137	0.012	(0.581, 0.661)
	37	0.784	-0.003	0.007	(0.755, 0.805)	-0.174	0.009	(0.583, 0.637)	-0.173	0.012	(0.574, 0.645)	-0.155	0.012	(0.587, 0.670)
Half	17	0.641	-0.001	0.010	(0.606, 0.675)	0.039	0.010	(0.639, 0.718)	0.030	0.012	(0.620, 0.718)	0.000	0.012	(0.598, 0.683)
	25	0.702	-0.002	0.006	(0.681, 0.721)	-0.020	0.006	(0.658, 0.706)	0.000	0.011	(0.667, 0.735)	0.006	0.012	(0.670, 0.745)
	33	0.758	-0.003	0.006	(0.735, 0.775)	-0.074	0.007	(0.663, 0.705)	-0.082	0.012	(0.641, 0.710)	-0.062	0.011	(0.658, 0.734)
	37	0.784	-0.003	0.007	(0.755, 0.804)	-0.098	0.009	(0.659, 0.712)	-0.097	0.011	(0.652, 0.721)	-0.077	0.011	(0.667, 0.747)

aBias: average bias, aSD: average standard deviation, CI: credible interval, aGA-con: regression model with continuous aGA, pGA-con: regression model with continuous pGA, pGA-cat: regression model with categorical pGA.

Table 3:

Simulation results when sample size is two times of that in the Scandinavian Study data.

Variability	Weeks	True AUC	aGA-con			pGA-con			pGA-cat			naïve Bi-Normal		
			aBias	aSD	95% CI	aBias	aSD	95% CI	aBias	aSD	95% CI	aBias	aSD	95% CI
Double	17	0.641	-0.002	0.007	(0.614, 0.662)	-0.085	0.008	(0.524, 0.587)	-0.088	0.009	(0.514, 0.591)	-0.101	0.009	(0.510, 0.569)
	25	0.702	-0.002	0.004	(0.685, 0.713)	-0.145	0.005	(0.539, 0.575)	-0.139	0.009	(0.532, 0.595)	-0.136	0.009	(0.534, 0.601)
	33	0.758	-0.003	0.004	(0.741, 0.769)	-0.200	0.005	(0.543, 0.573)	-0.203	0.009	(0.531, 0.579)	-0.192	0.009	(0.537, 0.595)
	37	0.784	-0.003	0.005	(0.764, 0.799)	-0.225	0.007	(0.539, 0.578)	-0.226	0.009	(0.532, 0.583)	-0.215	0.009	(0.538, 0.599)
Equal	17	0.641	-0.002	0.007	(0.613, 0.662)	-0.035	0.008	(0.577, 0.636)	-0.041	0.009	(0.562, 0.637)	-0.063	0.009	(0.548, 0.606)
	25	0.702	-0.002	0.004	(0.685, 0.713)	-0.094	0.005	(0.590, 0.625)	-0.082	0.009	(0.590, 0.650)	-0.076	0.009	(0.594, 0.658)
	33	0.758	-0.002	0.004	(0.741, 0.770)	-0.149	0.005	(0.595, 0.625)	-0.154	0.009	(0.580, 0.629)	-0.135	0.008	(0.594, 0.650)
	37	0.784	-0.003	0.005	(0.764, 0.798)	-0.173	0.007	(0.591, 0.631)	-0.174	0.009	(0.584, 0.636)	-0.155	0.009	(0.599, 0.658)
Half	17	0.641	-0.002	0.007	(0.614, 0.663)	0.038	0.007	(0.651, 0.705)	0.029	0.008	(0.636, 0.704)	-0.001	0.008	(0.611, 0.669)
	25	0.702	-0.002	0.004	(0.685, 0.713)	-0.020	0.004	(0.665, 0.698)	0.000	0.008	(0.676, 0.730)	0.007	0.008	(0.681, 0.738)
	33	0.758	-0.003	0.004	(0.741, 0.769)	-0.073	0.005	(0.670, 0.700)	-0.081	0.008	(0.653, 0.702)	-0.060	0.008	(0.671, 0.724)
	37	0.784	-0.003	0.005	(0.763, 0.798)	-0.098	0.006	(0.668, 0.705)	-0.097	0.008	(0.662, 0.711)	-0.076	0.008	(0.681, 0.735)

aBias: average bias, aSD: average standard deviation, CI: credible interval, aGA-con: regression model with continuous aGA, pGA-con: regression model with continuous pGA, pGA-cat: regression model with categorical pGA.

Table 4:

Simulation results when sample size is half of that in the Scandinavian Study data.

Variability	Weeks	True AUC	aGA-con			pGA-con			pGA-cat			naïve Bi-Normal		
			aBias	aSD	95% CI	aBias	aSD	95% CI	aBias	aSD	95% CI	aBias	aSD	95% CI
Double	17	0.641	-0.003	0.014	(0.589, 0.687)	-0.086	0.016	(0.489, 0.615)	-0.090	0.018	(0.474, 0.621)	-0.104	0.018	(0.479, 0.592)
	25	0.702	-0.003	0.008	(0.671, 0.731)	-0.146	0.010	(0.518, 0.594)	-0.138	0.018	(0.504, 0.622)	-0.136	0.019	(0.504, 0.628)
	33	0.758	-0.003	0.008	(0.726, 0.784)	-0.201	0.011	(0.526, 0.589)	-0.205	0.018	(0.507, 0.600)	-0.195	0.017	(0.509, 0.620)
	37	0.784	-0.003	0.010	(0.745, 0.815)	-0.226	0.014	(0.518, 0.598)	-0.227	0.018	(0.503, 0.608)	-0.216	0.018	(0.503, 0.628)
Equal	17	0.641	-0.003	0.014	(0.588, 0.687)	-0.037	0.015	(0.541, 0.663)	-0.044	0.017	(0.521, 0.667)	-0.066	0.017	(0.515, 0.630)
	25	0.702	-0.003	0.008	(0.671, 0.730)	-0.095	0.010	(0.570, 0.643)	-0.082	0.017	(0.563, 0.674)	-0.077	0.018	(0.565, 0.684)
	33	0.758	-0.003	0.008	(0.726, 0.784)	-0.149	0.011	(0.579, 0.641)	-0.155	0.017	(0.556, 0.649)	-0.138	0.017	(0.566, 0.674)
	37	0.784	-0.003	0.010	(0.745, 0.815)	-0.173	0.014	(0.573, 0.649)	-0.174	0.017	(0.558, 0.659)	-0.156	0.018	(0.566, 0.686)
Half	17	0.641	-0.003	0.014	(0.587, 0.688)	0.036	0.014	(0.620, 0.729)	0.026	0.016	(0.596, 0.730)	-0.003	0.017	(0.578, 0.693)
	25	0.702	-0.003	0.008	(0.671, 0.730)	-0.021	0.009	(0.646, 0.712)	0.000	0.016	(0.649, 0.752)	0.006	0.017	(0.653, 0.760)
	33	0.758	-0.003	0.008	(0.726, 0.784)	-0.074	0.010	(0.656, 0.714)	-0.083	0.016	(0.629, 0.721)	-0.063	0.016	(0.646, 0.746)
	37	0.784	-0.003	0.010	(0.743, 0.814)	-0.098	0.013	(0.650, 0.722)	-0.098	0.016	(0.635, 0.734)	-0.077	0.016	(0.649, 0.761)

aBias: average bias, aSD: average standard deviation, CI: credible interval, aGA-con: regression model with continuous aGA, pGA-con: regression model with continuous pGA, pGA-cat: regression model with categorical pGA.

Table 5:

Simulation results on efficiency gain when constraint is used.

Model	Mean	$aSD \times 10$
With constraint	0.670	0.201
Without constraint	0.673	0.236

Fitted models are based on equation (6). For “With constraint” model a truncated prior on β_1 was used, whereas for “Without constraint” model, unconstrained prior of β_1 was used. aSD: average standard deviation.

Table 6:

Model selection results using the Deviance Information Criterion (DIC), the Scandinavian Study.

Mean	Specification	Stage 1			Stage 2		
		\bar{D}	p_D	DIC	\bar{D}	p_D	DIC
Linear	$var(\epsilon_i) = \sigma_i^2$	-18972	2.940	-18969	1703	3.856	1706
Quadratic	$var(\epsilon_i) = \sigma_i^2$	-25096	4.139	-25092	1704	4.079	1708
Quadratic	$var(\epsilon_i) = \sigma_i^2 aGA_i^2$	-24069	4.109	-24065	1763	3.856	1767
Quadratic	$var(\epsilon_i) = \sigma_i^2 aGA_i^4$	361.7	4.151	365.8	2158	4.039	2162

DIC values with boldface are the lowest within its stage, and the corresponding models (under “Specifications”) are the best according to DIC. \bar{D} : posterior mean deviance, p_D : effective number of parameter.

Table 7:

Posterior AUC estimates, the Scandinavian Study data.

Models		Weeks of gestation					
		17	20	25	30	33	37
A	Mean	0.519	0.579	0.673	0.757	0.801	0.852
	$SD \times 10$	0.191	0.183	0.164	0.140	0.132	0.118
	95% CI	(0.483, 0.556)	(0.544, 0.614)	(0.642, 0.705)	(0.73, 0.784)	(0.776, 0.827)	(0.829, 0.874)
B	Mean	0.519	0.579	0.675	0.760	0.804	0.855
	$SD \times 10$	0.191	0.184	0.169	0.147	0.136	0.118
	95% CI	(0.483, 0.556)	(0.544, 0.614)	(0.641, 0.707)	(0.730, 0.788)	(0.777, 0.829)	(0.832, 0.878)
C	Mean	0.519	0.579	0.675	0.760	0.804	0.855
	$SD \times 10$	0.196	0.184	0.164	0.147	0.137	0.121
	95% CI	(0.481, 0.557)	(0.543, 0.615)	(0.642, 0.706)	(0.731, 0.788)	(0.777, 0.831)	(0.831, 0.878)
D	Mean	0.519	0.579	0.674	0.759	0.804	0.855
	$SD \times 10$	0.198	0.188	0.169	0.150	0.139	0.123
	95% CI	(0.481, 0.557)	(0.543, 0.615)	(0.642, 0.706)	(0.730, 0.788)	(0.777, 0.830)	(0.831, 0.878)
E	Mean	0.486	0.537	0.620	0.698	0.740	0.792
	$SD \times 10$	0.137	0.130	0.120	0.111	0.107	0.101
	95% CI	(0.460, 0.513)	(0.512, 0.562)	(0.597, 0.643)	(0.676, 0.719)	(0.720, 0.761)	(0.773, 0.811)
F	Mean	0.544		0.664		0.776	0.830
	$SD \times 10$	0.209	–	0.192	–	0.152	0.136
	95% CI	(0.504, 0.584)		(0.626, 0.700)		(0.746, 0.804)	(0.803, 0.855)

Model A: Unconstrained fixed effect Normal regression model using fixed PV;

Model B: Unconstrained random effect Normal regression model using fixed PV;

Model C: Constrained random effect Normal regression model using fixed PV;

Model D: Constrained random effect Normal regression model using varying PV;

Model E: Naïve regression model with continuous pGA as covariate and random effects and using varying PV;

Model F: Naïve regression model with pGA as categorical covariate with random effects and using varying PV.