Appendix 1

A two-stage modeling approach was used to evaluate biomarker change over time in relation to the outcomes of interest as a specified endpoint. In the first stage, the six annual biomarker values from each woman were assumed to be from the following linear mixed model

$$\log(\text{biomarker}_{ij}) = b_{i1} + b_{i2}t_{ij} + e_{ij} \quad i = 1, 2, ..., 50, \tag{1}$$

where t_{ij} is time since study, b_{i1} is the true biomarker value in log scale from woman i, b_{i2} is her annual change rate, and e_{ij} is the measurement error. The woman-specific variables b_{i1} and b_{i2} are assumed to be from the following bivariate normal distribution

$$\begin{bmatrix} b_{i1} \\ b_{i2} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, G = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \right)$$

and are independent of the measurement error e_{ij}, which is assumed to be independent and identically distributed as $\mathrm{N}(0, \sigma_e^2)$.

In the second stage, the relation between an endpoint of interest Y_i (such as age at the final menstrual period) and the woman-specific baseline value b_{i1} and change rate b_{i2} of a biomarker is evaluated in the following regression model

$$Y_i = \beta_0 + b_{i1}\beta_1 + b_{i2}\beta_2 + s_i^T\gamma_i + \epsilon_i \tag{2}$$

where β_1 is the effect of the true baseline biomarker on Y, β_2 is the effect of the change rate of the biomarker on Y, adjusting for other possible covariates s_i, and ϵ_i is the residual error.

However, the woman-specific variables b_{i1} and b_{i2} in model (2) are not observable. A common practice is to use the regression calibration approach where we replace b_{i1} and b_{i2} by their conditional means given the observed biomarker values in model (1) and then conduct regression analysis, which requires good estimates of the parameters μ_1, μ_2, G and σ_e^2. This approach works well when the measurement error σ_e^2 in model (1) is small to moderate (Wang, Wang and Wang, 2000; Li, Zhang and Davidian, 2004). Our problem is further complicated by the fact that some of the biomarker values are below the corresponding detection limits.

Thiebaut et. al (2004) indicated that the maximum likelihood estimation of model (1) can be implemented using the non-linear mixed model procedure Proc NLMixed of SAS to address the below-detection values. After the maximum likelihood estimates of these parameters are obtained, numerical integration can be used to calculate the conditional means of b_{i1} and b_{i2} for the regression calibration model

$$Y_i = \beta_0 + \widehat{b}_{i1}\beta_1 + \widehat{b}_{i2}\beta_2 + s_i^T\gamma_i + \epsilon_i^*, \tag{3}$$

where $\widehat{b}_{i1}$ and $\widehat{b}_{i2}$ are conditional means of b_{i1} and b_{i2} given the biomarkers in model (1).

After the woman-specific variables b_{i1} and b_{i2} are replaced by their conditional means, ϵ_i^* in model (3) tends to have unequal variances. Therefore, generalized estimating equation (GEE) approach as implemented in Proc Genmod of SAS, which only requires correct specification of the mean structure, is used to make inference of the regression parameters of interest in model (3).