Appendix 1

Analyses were implemented in Matlab7.0 (The MathWorks, Inc.), SAS version 9.1, SAS macro language and SAS/IML (SAS Institute, Cary, NC.).

A: Semi/non-parametric stochastic mixed model

Relationships between lumbar spine and femoral neck BMD and their simultaneous changes over time in relation to final menstrual period (FMP) could not be appropriately modeled by using quadratic or cubic terms; therefore, a semi-parametric stochastic mixed modeling was used. In general, the semi-parametric stochastic mixed model can be formulated by:

$$y_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + f(t_{ij}) + \boldsymbol{Z}_{ij}^T \boldsymbol{b}_i + U_i(t_{ij}) + \varepsilon_{ij} \quad \text{(A.1)}$$

where:

t_{ij} is the time variable (number of years prior or after FMP or FSH Stage) for subject i at j^{th} measurement;

$\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients associated with covariates $\boldsymbol{X}_{ij}$

$f(t)$ is a twice-differentiable smooth function of time;

$\boldsymbol{b}_i$ are independent $q \times 1$ vectors of random effects associated with covariates $\boldsymbol{Z}_{ij;}$

$U_i(t_{ij})$ are independent random processes used to model serial correlation;

ε_{ij} are independent measurement errors.

The fundamental assumptions for this model are: $\boldsymbol{b}_i$, $U_i(t_{ij})$, and ε_{ij} are mutually independent. $\boldsymbol{b}_i$ ~ normal ($\boldsymbol{0}$, $\boldsymbol{D}(\boldsymbol{\varphi})$), $\boldsymbol{D}$ is a positive definite matrix depending on a parameter vector $\boldsymbol{\varphi}$; $U_i(t_{ij})$ is a mean zero Gaussian process with covariance function or a non-homogeneous Ornstein-Uhlenbeck (NOU) process, cov($U_i(t)$,$U_i(s)$) = γ($\boldsymbol{\zeta}$, α; t, s) depending on a parameter vector $\boldsymbol{\zeta}$

and a scalar α, which is used to characterize the variance and correlation of the process $U_i(t)$;

ε_{ij} ~ iid $N(0,\sigma^2)$.

To capture the characteristics of BMD mean and variance varying over time, the modeling of BMD values (or rate of change) was formulated as:

$$y_{ij} = f(t_{ij}) + b_{0i} + b_{1i}t_{ij} + U_i(t_{ij}) + \varepsilon_{ij} \quad \text{(A.2)}$$

where $U_i(t)$ is a non-homogeneous Ornstein-Uhlenbeck (NOU) process satisfying:

$$\text{var}(U_i(t)) = \xi(t) \text{ with } \log(\xi(t)) = A_0 + A_1 t + A_2 t^2 \quad \text{(A.3)}$$

$$corr(U_i(t), U_i(s)) = \rho^{|t-s|} \quad \text{(A.4)}$$

This assumed each woman's serial correlation was the same. The smoothing function $f(t)$ represents the mean profile of BMD (or rate of change) for the population of women over time.

In this study, the potential covariates considered included body mass index (BMI), smoking behavior, hormones including FSH and the classifications of FSH into four stages and BMI into obese vs. non-obese.

B: **Subject-specific spline curves for longitudinal BMD data**

The subject-specific spline curves for the longitudinal BMD data were modeled as penalized splines with random effects within the mixed model framework. This approach allowed for a trade-off between spline regression and smoothing splines by relaxing the importance of the number of position of the knots and reducing the computational burden by using low-rank smoothers for large data sets.[1,2,3]

Let $\kappa_k (k = 1,2,\cdots,K)$ be the set of distinct knots in the time t_{ij} range and

$\left(t_{ij} - \kappa_k\right)_+ = \begin{cases} t_{ij} - \kappa_k & , t_{ij} > \kappa_k \\ 0 & , t_{ij} \le \kappa_k \end{cases}$ be the truncation function. The general form of subject-specific curves can be modeled using the penalized method:

$$y_{ij} = F(t_{ij}) + f_i(t_{ij}) + \varepsilon_{ij} \quad \text{(A.5)}$$

where:

$F(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \cdots + \beta_p t_{ij}^p + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+^p$ is a smooth function with p^{th} degree reflecting overall BMD change over time;

$f_i(t_{ij}) = b_0 + b_1 t_{ij} + \cdots + b_p t_{ij}^p + \sum_{k=1}^{K} v_{ik} (t_{ij} - \kappa_k)_+^p$ is the non-parametric subject-specific differences to the overall trend. A low-rank smoother with degree p = 1 or 2 can be used for large data sets.

$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ represents the measurement errors; $u_k \sim N(0, \sigma_u^2); v_{ik} \sim N(0, \sigma_v^2)$;

$(b_0, b_1, \cdots, b_p)^T \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$.

Model (A.9) was implemented using mixed models and by treating $(t_{ij} - \kappa_k)_+^p$ .and

$t_{ij}, \cdots, t_{ij}^p$ in $f_i(t_{ij})$ as random effects.

C: Change characteristics of BMD (or rate of change) trajectory

BMD changes (losses) over time follow a non-linear pattern. The instantaneous change of these trajectories (i.e., mean BMD profile) can be characterized by rate of change, acceleration / deceleration, and curvature. These can be approximated by the first- and second-order derivatives of the mean curve, and the hinge/bend of the mean curve integrating the rate of change and acceleration, respectively. The cubic spline approach was used to estimate the rate of change as well as acceleration or deceleration.

Assume the time *t* was equally spaced with step $h = t_{k+1} - t_k$ $(k = 1,2,...,n-1)$, where n was the total number of distinguishable time points. Let $f(t)$ be the BMD mean profile (trajectory) and $S_3(t)$ be its cubic spline approximation .The rate of change can be approximated by solving "m" equations:

$$m_{k-1} + 4m_k + m_{k+1} = 3\frac{f(t_{k+1}) - f(t_{k-1})}{h}, k = 2,3,...,n-1 \quad \text{(A.6)}$$

where m_k is the cubic spline approximation to $f'(t_k)$ with errors $O(h^4)$, $k = 2,3,...,n-1$. The m_1 and m_n can be given or computed by suitable forward and backward finite difference formulas respectively, e.g., using first 5-points and last 5-points,

$$m_1 = \frac{1}{12h}\left[-25f(t_1) + 48f(t_2) - 36f(t_3) + 16f(t_4) - 3f(t_5)\right] \quad \text{(A.7)}$$

$$m_n = \frac{1}{12h}\left[3f(t_{n-4}) - 16f(t_{n-3}) + 36f(t_{n-2}) - 48f(t_{n-1}) + 25f(t_n)\right] \quad \text{(A.8)}$$

The acceleration / deceleration can be approximated by solving "M" equations:

$$M_{k-1} + 4M_k + M_{k+1} = 6\frac{f(t_{k+1}) - 2f(t_k) + f(t_{k-1})}{h^2}, k = 2,3,...,n-1 \quad \text{(A.9)}$$

where M_k is the second derivative approximations $f''(t_k)$ with $O(h^2)$ errors, $k = 2,3,...,n-1$.

The M_1 and M_n satisfied the boundary conditions $2M_1 + M_2 = \frac{6}{h}\left(\frac{f(t_2) - f(t_1)}{h} - m_1\right)$ and

$M_{n-1} + 2M_n = \frac{6}{h}\left(m_n - \frac{f(t_n) - f(t_{n-1})}{h}\right)$.

The population rate of BMD change and 95% CI over time around the FMP was obtained by using non-parametric stochastic mixed modeling, or bootstrapping 100 samples.
The associations between rate of BMD change and BMI, FSH stages and/or hormones (including FSH, E2) were modeled using semi-parametric stochastic mixed models.

D: **Piecewise linear mixed model related time (ovarian aging and chronological aging)** (4)

The nodes (or turning points) of the population BMD profile were identified based on the changing characteristics obtained by the above processes and they were further used to segment the hormone trajectory into stages. Then, piecewise linear mixed models were

developed to identify segment characteristics (i.e., rate of change for each segment). Statistical comparisons of these slopes from two consecutive intervals around a turning point were tested to ascertain if one slope was different than the adjacent slope. The piece wise linear mixed model was formulated as follows.

Assume the independent time variable of interest $t \in T \subset R^1$ (e.g., the time to FMPor time in FSH Stages). Let $\Omega = \left\{ t^*_{(k)}, 1 \le k \le K \middle| t^*_{(1)} < t^*_{(2)} < \ldots < t^*_{(K)} \right\}$ be one known division of T, where K is the total number of turning points used to split T into $(K+1)$ non-overlapped intervals. The mean structure of piecewise linear mixed effect model was given by:

$$E[y_{ij}] = \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} \beta_k^{(1)} \left(t_{ij} - t^*_{(k)} \right)_+ + \sum_{l=1}^{L} \beta_l^{(2)} x_{ijl} + \sum_{l=1}^{L} \sum_{k=1}^{K} \beta_{l,k} \left(t_{ij} - t^*_{(k)} \right)_+ x_{ijl} \qquad \text{(A.10)}$$

where $\left(t_{ij} - t^*_{(k)} \right)_+ = \begin{cases} t_{ij} - t^*_{(k)} & , t_{ij} > t^*_{(k)} \\ 0 & , t_{ij} \le t^*_{(k)} \end{cases}$; $x_{ijl} (l = 1,2,..., L)$ are the covariates of interest. If there are no other covariates (e.g., pure "time" effect is of interest) then $\beta_l^{(2)}$ and $\beta_{l,k}$ can be dropped from the model. The random effects will be appropriately specified, e.g., random intercept and random slopes. The variance-covariance structure and model assumptions follow those of general linear mixed models.

REFERENCES

1. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. 1st ed. Cambridge, UK: Cambridge University Press; 2003.
2. Durbán M, Harezlak J, Wand MP, Carroll RJ. Simple fitting of subject-specific curves for longitudinal data. *Stat Med*. 2005;24(8):1153-1167.
3. **Zhang D, Lin X, Sowers M** 1998 Semiparametric stochastic mixed models for longitudinal data. J Am Stat Assoc 93:710-719
4. **Efron B, Tibshirani R** 1986 Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Science 1:54-77
5. **Claeskens G, Van Keilegom I** 2003 Bootstrap confidence bands for regression curves and their derivatives. Ann Statistics 31:1852-1884
6. **Neter J, Wasserman W, Kutner M** 1985 Applied linear statistical models. 2nd ed. Homewood, Illinois: Irwin