Methodological Aspects of Studying Human Aging, Health and Mortality

AI Yashin, I Akushevich, K Arbeev, A Kulminski, S Ukraintseva,
Center for Population Health and Aging, Duke University, USA

Introduction

Traditional (demographic) description of changes in individual health status is performed using continuous time random Markov process with finite number of states, and age dependent transition intensity functions (transitions rates). Although such a description of the health process plays an important role in understanding connection between health and mortality it did not allow for studying factors and mechanisms involved in aging related decline in health/well-being/survival status. Numerous epidemiological studies provide compelling evidence that health transitions rates are influenced by a number of variables. Some of them are fixed at the time of birth (e.g., genetic background). Others experience stochastic changes over the life course. The latter include physiological state, medical cost, behavioral, or social-economical factors, etc. The presence of such randomly changing influential factors violates Markov assumption, and makes the description of aging related changes in health status more complicated.

The dynamics of such influential factors (e.g. physiological variables) in connection with mortality risks has been described using stochastic process model of human mortality and aging (Woodbury and Manton, 1977). Recent extensions of this model have been used in the analyses of longitudinal data on aging, health, and longevity, collected in the Framingham Heart Study (Yashin et al., 2007; 2008; Arbeev at al., 2009). This model and its extensions enjoyed Markov property of stochastic process satisfying diffusion type stochastic differential equation. The stochastic process is stopped at random time associated with individual's death. The quadratic hazard assumption about the form of conditional mortality, given covariates values and certain regularity conditions, guarantee Gaussian property of conditional distribution of the covariates values at any given age. This allowed for description of aging related changes in terms of two first moments of multidimensional Gaussian distribution. When individual's health status is taken into account the coefficients of stochastic differential equations become dependent from values of jumping process. This dependence violates Markov assumption and makes conditional Gaussian property not valid. So the description of this (continuously changing) component of aging related changes in the body also becomes more complicated.

Since studying age trajectories of physiological state in connection with changes in health status would provide more realistic scenario for analyses of available longitudinal data it would be a good idea to find an appropriate description of these two interdependent processes developing in aging organism. For this purpose we suggest a comprehensive model of human aging, health and mortality where Markov assumption is fulfilled for the stochastic process consisting of jumping and continuously changing components. The jumping component is used for description of relatively fast changes in health status, and continuous component describes relatively slow age-dynamics of individual physiological state.

The Model

Let $\theta_t, t \geq 0$ be the finite-state (jumping) stochastic process (i.e., $\theta_t \in \{1,2,...,M\}$, where M is the number of states) and $Y_t, t \geq 0$ be $K-$dimensional stochastic process with continuous components describing joint evolution of individual health/well-being status and physiological variables over age. We assume that Y_t satisfies a stochastic differential equation with coefficients depending on θ_t:

$$dY_t = A_{\theta_t}(Y_t, t)dt + B_{\theta_t}(t)dW_t, Y_{t_0} \quad (1)$$

Here $A_{\theta_t}(Y_t, t)$ is a vector function, $B_{\theta_t}(t)$ is a matrix of respective dimension, Y_{t_0} is a random vector of initial conditions, and W_t is a vector Wiener process with independent components which is

independent from initial value, Y_{t_0}.

The finite state continuous time process θ_t, describing jumping changes in health/well-being status is characterized by conditional transition intensity matrix $\lambda_{kr}(Y_t,t), k,r=1,2,...,M; \lambda_{kk}(Y_t,t)=-\sum_{r=1,r\neq k}^{M}\lambda_{kr}(Y_t,t)$, and initial probabilities $P(\theta_{t_0}=j)$, *j=1,2,…M.*

Let T be non-negative random variable, describing life span. Its distribution characterizes variability in life span among individuals in human cohorts representing longitudinal data. Individual's death at time T means that the trajectories of θ_t and Y_t are stopped at time T. The conditional distribution of T given trajectories of θ_u, $Y_u, 0\le u\le t$ is completely characterized by the conditional hazard (mortality) rate $\mu_{\theta_t}(Y_t,t)$.

The use of stochastic differential equations for random continuously changing covariates has been studied intensively in the analysis of longitudinal data (Yashin et al., 2007; 2008; Arbeev et al. 2009, and references therein). Such description is convenient since it captures a feedback mechanism typical of biological systems reflecting regular aging related changes and takes into account the presence of random noise affecting individual trajectories. It also captures dynamic connection between health and physiological states, which is important in many applications.

Survival analysis using model of human aging, health, and mortality

Let $T_1,T_2,...,T_N$ be life span data on N individuals, which health status and physiological state are described by the processes θ_t and Y_t. The likelihood function of these data is:

$$L(T_1,T_2,...T_N)=\prod_{i=1}^{N}\bar{\bar{\mu}}(T_i)^{\delta_i}\exp\left\{-\int_0^{T_i}\bar{\bar{\mu}}(u)du\right\} \qquad (2)$$

Here $\bar{\bar{\mu}}(t)=E\left(\mu_{\theta_t}(Y_t,t)\middle|T>t\right)$, and δ_i is censoring variable. The likelihood function (2) has to be maximized with respect to parameters describing total mortality rate, $\bar{\bar{\mu}}(t)$. Since these parameters are involved in characterization of the process θ_t, Y_t and probability distribution of T, their interpretation has biological and physiological sense.

Nonlinear partial differential equation for conditional p.d.f./probability.

To calculate $\bar{\bar{\mu}}(t)$ one needs $f(y,j|t)=\frac{\partial}{\partial y}P\left(Y_t\le y,\theta_t=j\middle|T>t\right)$ which is the joint conditional probability density function, p.d.f. with respect to Y_t, and the probability with respect to θ_t, given $\{T>t\}$**.** Using standard Bayesian arguments similar to that used in Yashin et al. (1985, 1994) the following partial differential equation for this function can be derived:

$$\frac{d}{dt}f(y,j|t)=\sum_{i=1}^{M}\lambda_{ij}(y,t)f(y,i|t)-\frac{\partial}{\partial y}\left(A_j(y,t)f(y,j|t)\right)+$$
$$+\frac{1}{2}\frac{\partial^2}{\partial y^2}\left(B_j(t)f(y,j|t)\right)+f(y,j|t)\left(\bar{\bar{\mu}}(t)-\mu_j(y,t)\right),\quad f(y,j|t_0) \qquad (3)$$

Here functions $A_j(y,t)$, and $B_j(t)$ are defined in (1). Since $f(y,j|t)$ multiplies $\bar{\bar{\mu}}(t)$ in (2), this equation is nonlinear partial differential equation with respect to $f(y,j|t)$. The total mortality rate $\bar{\bar{\mu}}(t)$ can also be represented as follows:

$$\bar{\bar{\mu}}(t) = \sum_{j=1}^{M} \bar{\mu}_j(t)\pi_j(t) \tag{4}$$

where $\pi_j(t) = P(\theta_t = j \,|\, T > t)$, and

$$\bar{\mu}_j(t) = \sum_{j=1}^{M} E\left(\mu_{\theta_t}(Y_t,t) \,|\, \theta_t = j, T > t\right) \tag{5}$$

To calculate (4) and (5) one needs $\pi_j(t) = P(\theta_t = j \,|\, T > t)$ and conditional p.d.f., $f(y \,|\, j,t) = \partial P\left(Y_t \le y \,|\, \theta_t = j, T > t\right)/\partial y$ for each $t \ge 0$. Equation for $\pi_j(t)$ can be derived by integrating $f(y, j \,|\, t)$ in (3) with respect to y:

$$d\pi_j(t)/dt = \sum_{k=1}^{M} \bar{\lambda}_{k,j}(t)\pi_k(t) + \pi_j(t)\left(\bar{\bar{\mu}}(t) - \bar{\mu}_j(t)\right), \pi_j(t_0) \quad j = 1,2,\ldots M \tag{6}$$

Here $\bar{\bar{\mu}}(t)$ and $\bar{\mu}_j(t)$ are given by (4) and (5), and $\bar{\lambda}_{ij}(t)$ is defined as follows:

$$\bar{\lambda}_{ij}(t) = E\left(\lambda_{ij}(Y_t,t) \,|\, \theta_t = i, T > t\right) = \int_{R^n} \lambda_{ij}(y,t) f(y \,|\, i,t)\, dy. \tag{7}$$

Integration in (5) and (7) requires $f(y \,|\, j,t) = \partial P\left(Y_t \le y \,|\, \theta_t = j, T > t\right)/\partial y, ^j = 1,2,\ldots,M$.

Equation for $f(y \,|\, j,t) = \partial P\left(Y_t \le y \,|\, \theta_t = j, T > t\right)/\partial y$ follows from equations (3) and (6):

$$\frac{\partial}{\partial t} f(y \,|\, j,t) = \sum_{i=1}^{N} \left(\lambda_{ij}(y,t) f(y \,|\, i,t) - \bar{\lambda}_{ij}(t) f(y \,|\, j,t)\right) \frac{\pi_i(t)}{\pi_j(t)} - \frac{\partial}{\partial y}\left(A_j(y,t) f(y \,|\, j,t)\right)$$

$$+ \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(B_j(t) f(y \,|\, j,t)\right) + f(y \,|\, j,t)\left(\bar{\mu}_j(t) - \mu_j(y,t)\right), \quad f(y \,|\, j,t_0) \tag{8}$$

Note that (6) and (8) is a system of nonlinear (partial and ordinary) differential equations.

Gaussian Approximation

To solve equations (7) and (9) the functional forms for the coefficient, $A_{\theta_t}(Y_t,t)$ in (1) and (3) the elements of conditional transition intensities matrix $\lambda_{k,r}(Y_t,t)$, as well as for conditional mortality rate $\mu_{\theta_t}(Y_t,t)$ have to be specified, and respective integrations have to be performed to get $\bar{\bar{\mu}}(t)$, $\bar{\mu}_j(t)$ and $\bar{\lambda}_{ij}(t)$. It is convenient and epidemiologically justified to describe such functions as quadratic forms of variable Y_t:

$$\lambda_{kr}(Y,t) = \lambda_{0kr}(t) + \left(Y_t - g_k(t)\right)^* \Lambda_{kr}(t)\left(Y_t - g_k(t)\right) \tag{9}$$

$$\mu_{\theta_t}(Y_t,t) = \mu_{0\theta}(t) + \left(Y_t - f_{\theta_t}(t)\right)^* Q_{\theta_t}(t)\left(Y_t - f_{\theta_t}(t)\right) \tag{10}$$

Here $\Lambda_{kr}(t)$ and $Q_j(t)$, are symmetric non-negative-definite K x K matrices, $f_{\theta_t}(t)$ is a K-vector functions $\lambda_{0kr}(t)$ and $\mu_{0r}(t)$ are parametric functions of t for *k, r, j* = 1, 2,…,*M*; $t \ge t_0$. The includes negative feedback loops, which allow for maintaining organisms' functioning. It is convenient to describe mechanism of physiological regulation in the presence of external disturbances in terms of linear stochastic differential equation:

$$dY_t = a_{\theta_t}(t)\left(Y_t - f_{1\theta_t}(t)\right)dt + B_{\theta_t}(t)dW_t, Y_{t_0} \tag{11}$$

Here $a_{\theta_t}(t)$ is a vector function, $B_{\theta_t}(t)$ is a matrix of respective dimension, Y_0 is a random vector of initial conditions, and W_t is a vector Wiener process with independent components, which is independent from initial value, Y_0. The components of vector function $f_{1\theta_t}(t)$ characterize the effects of allostatic adaptation on physiological state (Yashin et al, 2007; 2008).

Conditions (10) and (11) together with the assumptions about normality of the distribution for Y_{t_0} guaranteed Gaussian property of conditional probability distribution of the process Y_t among survivors in the absence of jumping process (Yashin 1985; Yashin and Manton, 1997). The presence of jumping process θ_t affecting the structure of the equation (11) for Y_t, and hence its age dynamics violates Gaussian property. However, the quadratic forms for conditional transition intensity functions and mortality rates and linear structure of (11) suggests the possibility for using Gaussian approximation of the conditional p.d.f. $f(y|j,t) = \partial P\left(Y_t \le y | \theta_t = j, T > t\right)/\partial y$.

The conditional hazard (mortality rate) given health status $\theta_t = j,$ and unconditional transition intensity functions can be represented as follows:

$$\overline{\mu}_j(t) = \mu_{0j}(t) + \left(m_j(t) - f_j(t)\right)^* Q_j(t)\left(m_j(t) - f_j(t)\right) + Tr\left(Q_j(t)\gamma_j(t)\right) \quad (12)$$

$$\overline{\lambda}_{jk}(t) = \lambda_{0jk}(t) + \left(m_j(t) - g_j(t)\right)^* \Lambda_{jk}(t)\left(m_j(t) - g_j(t)\right) + Tr\left(\Lambda_{jk}(t)\gamma_j(t)\right) \quad (13)$$

where $m_j(t) = E\left(Y_t | \theta_t = j, T > t\right)$, and $\gamma_j(t) = E\left(\left(Y_t - m_j(t)\right)^*\left(Y_t - m_j(t)\right) | \theta_t = j, T > t\right)$.

These conditional moments satisfy the following ordinary differential equations (for brevity we omit dependence from t in all variables below):

$$\frac{dm_j}{dt} = \sum_i \frac{\pi_i}{\pi_j}\left[m_{ij}\overline{\lambda}_{ij} - 2\gamma_i\Lambda_{ij}\hat{g}_i\right] - a_j\hat{f}_{1j} + 2\gamma_j Q_j \hat{f}_j, \quad (14)$$

$$\frac{d\gamma_j}{dt} = \sum_i \frac{\pi_i}{\pi_j}\left[\left(\gamma_i - \gamma_j + m_{ij}\cdot m_{ij}^*\right)\overline{\lambda}_{ij} + 2\left(\gamma_i\Lambda_{ij}\gamma_i - \gamma_i\Lambda_{ij}\hat{g}_i\cdot m_{ij}^* - m_{ij}\cdot\hat{g}_i^*\Lambda_{ij}\gamma_i\right)\right]$$

$$+ a_j\gamma_j + \gamma_j a_j^* + B_j - 2\gamma_j Q_j \gamma_j. \quad (15)$$

Here $\overline{\lambda}_{ij}(t)$ is given by (14), $m_{ij} = m_i - m_j$, "hat" variables are defined as $\hat{f}_j = f_j - m_j$, $\hat{f}_{1j} = f_{1j} - m_j$, $\hat{g}_i = g_i - m_i$.

Note that the likelihood function (2) becomes a function of parameters determining dynamic properties of equations (6), (14), and (15). The fact that all these parameters have clear biological interpretation is an important advantage of our model compared to other parametric models of mortality used in demographic applications. There is a price for having proper interpretation, however: the equations (6), (14), (15) do not have explicit analytical solution. Therefore they have to be solved numerically at each step of the likelihood maximization procedure.

Observational plans

#1. When continuous variables are observed in discrete times

Let us assume now that continuously changing variables are measured at age points $t_0, t_1, t_2 ..., t_n; t_n < t \le T$. Let $\tilde{Y}_0^t = Y_{t_0}, Y_{t_1}, Y_{t_2}, ..., Y_{t_n}; t_n < t \le T$ be a random vector of observations of Y_t at these age points. It follows from these notations that $\tilde{Y}_0^{t_k-} = \tilde{Y}_0^{t_{k-1}}$ and $\tilde{Y}_0^t = \tilde{Y}_0^{t_k}$, if $t_k \le t < t_{k+1}$. Here $t_k - = \lim_{u \uparrow t_k} t_u$. Denote by

$$\tilde{\pi}_j(t) = P(\theta_t = j | \tilde{Y}_0^t, T > t)) \quad (16)$$

the conditional probability of having health/well-being status j, given $\tilde{Y}_0^t$, $\{T > t\}$. Let

$$\tilde{f}(y|j,t) = \frac{\partial}{\partial y} P\left(Y_t \le y \middle| \tilde{Y}_0^t, \theta_t = j, T > t\right) \quad (17)$$

The evolution of $\tilde{\pi}_j(t)$ and $\tilde{f}(y|j,t) = \frac{\partial}{\partial y} P\left(Y_t \le y \middle| \tilde{Y}_0^t, \theta_t = j, T > t\right)$ starts at age t_0, and continues at the intervals $t_0 \le t < t_1; t_1 \le t < t_2; ...; t_{n-1} \le t < t_n; t < T$. At each such interval these functions satisfy the equations (6) and (8).

An important property of the age trajectories of $\tilde{\pi}_j(t)$ and $\tilde{f}(y|j,t)$ is that they both will experience jumps at the observation times $t_1, t_2, ...t_n, t_n < t \le T$:

$$\tilde{\pi}_j(t_k) = \tilde{\pi}_j(t_k -) \frac{\tilde{f}(Y_{t_k}|j,t_k -)}{\sum_{r=1}^{M} \tilde{\pi}_r(t_k -) \tilde{f}(Y_{t_k}|r,t_k -)}; \quad \tilde{f}(y|j,t_k) = \delta(y - Y_{t_k}) \quad (18)$$

respectively. Here $\tilde{f}(Y_{t_k}|j,t_k -) = \frac{\partial}{\partial y} P\left(Y_t \le y \middle| \tilde{Y}_0^{t-}, \theta_t = j, T > t\right)_{t=t_k, y=Y_{t_k}}$, i.e. it is a solution of equation (9) at the interval $[t_{k-1}, t_k)$, taken at time point t_k, where y is replaced by observed value Y_{t_k}. In other words this function characterizes how likely it is to observe value Y_{t_k} under given conditions at time point t_k. The notations for transition coefficients and mortality rate under the new conditions corresponding to observational plan #1 are $\tilde{\tilde{\lambda}}_{k,j}(\tilde{Y}_0^t, t) = E\left(\lambda_{k,j}(Y_t, t) \middle| \tilde{Y}_0^t, \theta_t = k, T > t\right)$; $\tilde{\tilde{\mu}}(\tilde{Y}_0^t, t) = \sum_{j=1}^{M} \tilde{\mu}_j(\tilde{Y}_0^t, t) \tilde{\pi}_j(t)$; and $\tilde{\mu}_j(\tilde{Y}_0^t, t) = E\left(\mu(\theta_t, Y_t, t) \middle| \tilde{Y}_0^t, \theta_t = j, T > t\right)$. The notations, specified above, allow us to form the likelihood functions for the data on the sequence of discrete-time measurements of continuously changing component, plus survival data. The first part of likelihood requires p.d.f. $\tilde{\phi}\left(y|t\right) = \partial P\left(Y_t \le y \middle| \tilde{Y}_0^t, T > t\right)/\partial y$ which may be obtained as $\tilde{\phi}\left(y|t\right) = \sum_{r=1}^{M} \tilde{f}\left(y|r,t\right) \pi_r(t)$. Thus, the component of the likelihood function for the *i-th* individual having measurements $y_{t_1^i}^i, y_{t_2^i}^i, ...y_{t_{n(i)}^i}^i, T_i$ is:

$$L_i(y_{t_1^i}^i, y_{t_2^i}^i, ...y_{t_{n(i)}^i}^i, T_i) = \quad (19)$$

$$= \tilde{\phi}\left(y_{t_{n(i)}^i} \middle| t_{n(i)}^i -\right) \tilde{\phi}\left(y_{t_{n(i)-1}^i} \middle| t_{n(i)-1}^i -\right) ..... \tilde{\phi}\left(y_{t_1^i} \middle| t_1^i -\right) \tilde{\tilde{\mu}}^i(T_i)^{\delta_i} \exp\left\{-\int_0^{T_i} \tilde{\tilde{\mu}}^i(u) du\right\}$$

Here $\tilde{\phi}\left(y_{t_k^i} \middle| t_k^i -\right) = \sum_{r=1}^{M} \tilde{f}\left(y_{t_k^i} \middle| r, t_k^i -\right) \pi_r(t_k^i -)$.

Gaussian approximation of $\tilde{f}(y|j,t)$ in case of observational plan #1. In case of quadratic mortality risk, quadratic transition intensity functions, and linear equations for Y_t the likelihood function (19) can be represented as follows:

$$L_i\left(y_{t_0^i}^i, y_{t_1^i}^i, ..., y_{t_{n_i}^i}^i, T_i\right) = \tilde{\tilde{\mu}}^i(T_i)^{\delta_i} \exp\left\{-\int_0^{T_i} \tilde{\tilde{\mu}}^i(u) du\right\} \prod_{j=0}^{n_i(T_i)} \left[\sum_{k=1}^{M} \tilde{\pi}_k^i(t_j^i -)\right.$$

$$\left.\left(2\pi \left|\tilde{\gamma}_k^i(t_{t_j^i}^i -)\right|\right)^{-\frac{K}{2}} \exp\{-\frac{1}{2}\left(y_{t_j^i}^i - \tilde{m}_k^i(t_j^i -)\right)^* \tilde{\gamma}_k^i(t_{t_j^i}^i -)^{-1} \left(y_{t_j^i}^i - \tilde{m}_k^i(t_j^i -)\right)\right] \quad (20)$$

where $\tilde{\tilde{\mu}}(\tilde{Y}_0^t, t)$ is defined above, and $\tilde{\mu}_j(\tilde{Y}_0^t, t)$ is represented as:

$$\tilde{\mu}_j(\tilde{Y}_0^t, t) = \mu_{0j}(t) + \left(\tilde{m}_j(t) - f_j(t)\right)^* Q_j(t) \left(\tilde{m}_j(t) - f_j(t)\right) + Tr\left(Q_j(t) \tilde{\gamma}_j(t)\right) \quad (21)$$

$$\tilde{m}_k(t) = E\left(Y_t \middle| \tilde{Y}_0^t, \theta_t = k, T > t\right); \quad \tilde{\gamma}_k(t) = E\left(\left(Y_t - \tilde{m}_k(t)\right)^*\left(Y_t - \tilde{m}_k(t)\right)\middle| \tilde{Y}_0^t, \theta_t = k, T > t\right)$$

The transition intensities $\tilde{\lambda}_{kj}(\tilde{Y}_0^t, t)$ in equations (6) for $\tilde{\pi}_j(t)$ are:

$$\tilde{\lambda}_{kj}(\tilde{Y}_0^t, t) = \lambda_{0kj}(t) + \left(\tilde{m}_k(t) - g_k(t)\right)^* \Lambda_{kj}(t)\left(\tilde{m}_k(t) - g_k(t)\right) + Tr\left(\Lambda_{kj}(t)\tilde{\gamma}_k(t)\right) \tag{22}$$

Note that using index i in $\tilde{\pi}_k^i$, $\tilde{m}^i$ and $\tilde{\gamma}^i$ in these equations is needed because the values of these estimates depend on individual histories of the process Y_t observed in discrete times. Here δ_i is a censoring indicator, K is the dimension of vector Y_t, $m^i(t)$ and $\gamma^i(t)$ satisfy equations (14) and (15) at the intervals $\left[t_0^i, t_1^i\right)$; $\left[t_1^i, t_2^i\right)$;...;$\left[t_{n_i-1}^i, t_{n_i}^i\right)$; $\left[t_{n_i}^i, T_i\right)$ with the initial conditions $y_{t_0^i}^i, y_{t_1^i}^i, ..., y_{t_{n(i)}^i}^i$, respectively, $\tilde{m}^i(t_j^i-) = \lim_{t\uparrow t_j^i} \tilde{m}^i(t)$, and $\tilde{\gamma}^i(t_j^i-) = \lim_{t\uparrow t_j^i} \tilde{\gamma}^i(t)$, and $t_{n(i)}^i$ is the age of the latest measurement of the physiological index before death at T_i.

The conditions (18) can now be represented in the form:

$$\tilde{\pi}_j(t_i) = \tilde{\pi}_j(t_i-)\frac{\left(2\pi\left|\tilde{\gamma}_j(t_i-)\right|\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left(Y_{t_i} - \tilde{m}_j(t_i-)\right)^*\tilde{\gamma}_j^{-1}(t_i-)\left(Y_{t_i} - \tilde{m}_j(t_i-)\right)\right\}}{\sum_{k=1}^{M}\tilde{\pi}_k(t_i-)\left(2\pi\left|\tilde{\gamma}_k(t_i-)\right|\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left(Y_{t_i} - \tilde{m}_k(t_i-)\right)^*\tilde{\gamma}_k^{-1}(t_i-)\left(Y_{t_i} - \tilde{m}_k(t_i-)\right)\right\}},$$

$\tilde{m}_j(t_i) = Y_{t_i}$ and $\tilde{\gamma}_j(t_i) = 0$. The dynamics of $\tilde{\pi}_j(t)$ follows the equation (6) with $\tilde{\lambda}_{kj}(\tilde{Y}_0^t, t)$ used instead of $\overline{\lambda}_{kj}(t)$ at the intervals $t_0 \le t < t_1; t_1 \le t < t_2; ...; t_{n-1} \le t < t_n; t < T$. Note that the initial values of $\tilde{\pi}_j(t)$ at the beginning of i^{th} interval $[t_i \le t < t_{i+1})$ are given by the relationship which involves values of $\tilde{\pi}_j(t_i-)$, $\tilde{m}_j(t_i-)$ and $\tilde{\gamma}_j(t_i-)$ which are the solution of the equations (6), (14), and (15) at the end of the interval $[t_{i-1} \le t < t_i)$.

#2. Measuring changes in health state. No measurements of physiological state

Let $\hat{f}(y,t) = \frac{\partial}{\partial y}P\left(Y_t \ge y \middle| \theta_0^t, T > t\right)$, and $\tau_1, \tau_2, ..., \tau_m$ are ages at which changes in person's health status took place (time moments of jumps of the process θ_t). Then for conditional probability density function of Y_t given age trajectories of health history θ_0^t, and $\{T > t\}$, the following equation takes place:

$$\frac{\partial}{\partial t}\hat{f}(y,t) = -\frac{\partial}{\partial y}\left(A_{\theta_t}(y,t)\hat{f}(y,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(B_{\theta_t}(t)\hat{f}(y,t)\right) + \tag{23}$$

$$+\hat{f}(y,t)\left(\sum_{k=1,k\ne\theta_{t-}}^{M}\tilde{\lambda}_{\theta_{t-},k}(t) - \sum_{k=1k\ne\theta_{t-}}^{M}\lambda_{\theta_{t-},k}(y,t)\right) + \hat{f}(y,t)\left(\tilde{\mu}_{\theta_{t-}}(t) - \mu_{\theta_{t-}}(y,t)\right)$$

This equation has to be solved at the intervals $[\tau_1, \tau_2), [\tau_2, \tau_3), ..., [\tau_m, T)$, i.e., between subsequent jumps of the process θ_t. To avoid multiple hierarchical indexing we will use notation $\theta_t \equiv \theta(t)$. The initial conditions at the beginning of each interval are

$$\tilde{f}_{\tau_p}(y) = \tilde{f}_{\tau_p-}(y)\frac{\lambda_{\theta(\tau_p-),\theta(\tau_p)}(y,\tau_p)}{\hat{\lambda}_{\theta(\tau_p-),\theta(\tau_p)}(\tau_p)} \tag{24}$$

Here

$$\hat{\lambda}_{\theta(\tau_p-),\theta(\tau_p)}(\tau_p) = E\left(\lambda_{\theta(\tau_p-),x}(Y_t,\tau_p)\middle|\theta_0^{\tau_p-}, T > \tau_p\right)\Big|_{x=\theta(\tau_p)} \tag{25}$$

The likelihood function of the data on health transitions: observational plan #2. The part of the likelihood, corresponding to the Medicare data on ages of change in the health status (age at onset of diseases) for the *i-th* individual with $m(i)$ changes in the health status happened at the time points $\tau_1^i, \tau_2^i, ...\tau_{m(i)}^i$, and death (censoring) happened at age T_i is:

$$L_i(\tau_1^i, \tau_2^i, ...\tau_{m(i)}^i, T_i) = p(\theta^i(0)) \prod_{p=0}^{m(i)} \widehat{\lambda}_{\theta^i(\tau_p^i-),\theta^i(\tau_p^i)}(t_p^i) \exp\left\{ -\int_{t_{p-1}^i}^{t_p^i} \left(\sum_{k=1,k\neq\theta^i(t-)}^{M} \widehat{\lambda}_{\theta^i(t-),k}(t) + \right.\right.$$

$$\left.\left. + \widehat{\mu}_{\theta^i(t-)}(t) \right) dt \right\} \times \widehat{\mu}_{\theta^i(\tau_{m(i)}^i)}(T_i)^{\delta_i} \exp\left\{ -\int_{t_{m(i)}^i}^{T_i} \left(\sum_{k=1,k\neq\theta^i(t-)}^{M} \widehat{\lambda}_{\theta^i(t-),k}(t) + \widehat{\mu}_{\theta^i(t-)}(t) \right) dt \right\} \qquad (26)$$

Here $p(\theta^i(0))$ is the initial distribution of the health status, and $\tau_0^i = 0$ by definition.

Gaussian approximation in case of observational plan #2. Since health transitions are observed equations for the first two moments are simplified ($j = \theta_{t-}$).

$$\frac{dm}{dt} = -a\hat{f}_1 + \sum_{k\neq j} 2\gamma\Lambda_{jk}\hat{g}_j + 2\gamma Q_j \hat{f}, \qquad (27)$$

$$\frac{d\gamma}{dt} = a\gamma + \gamma a^* + B - \sum_{k\neq j} 2\gamma\Lambda_{jk}\gamma - 2\gamma Q_j \gamma. \qquad (28)$$

These equations has to be solved at the intervals $[\tau_1, \tau_2), [\tau_2, \tau_3), ..., [\tau_m, T)$, i.e., between subsequent jumps of the process θ_t. When $\theta(\tau_p-) = k$ and $\theta^i(\tau_p) = j$, and $\lambda_{kj}(Y,t)$ is described by (10), we have for the initial value $m(\tau_p)$:

$$m(\tau_p) = \frac{m(\tau_p-)\left[\lambda_{0kj}(\tau_p) + Tr(\gamma(\tau_p-)\Lambda_{kj}(\tau_p)) + \widehat{g}_k(\tau_p-)^*\Lambda_{kj}(\tau_p)\widehat{g}_k(\tau_p-)\right] - 2\gamma(\tau_p-)\Lambda_{kj}(\tau_p)\hat{g}_k(\tau_p-)}{\tilde{\lambda}_{kj}(\tau_p-)} \qquad (29)$$

$$\gamma(\tau_p) = \gamma(\tau_p-) \frac{\left[\lambda_{0kj}(\tau_p) + Tr(\gamma(\tau_p-)\Lambda_{kj}(\tau_p)) + \widehat{g}_k(\tau_p-)^*\Lambda_{kj}(\tau_p)\widehat{g}_k(\tau_p-)\right] + 2\Lambda_{kj}(\tau_p)\gamma(\tau_p-)}{\tilde{\lambda}_{kj}(\tau_p-)} \qquad (30)$$

With

$$\widehat{g}_k(\tau_p-) = g_k(\tau_p) - m(\tau_p-)$$

and

$$\tilde{\lambda}_{kj}(t) = \lambda_{0kj}(t) + \left(m_k(t) - g_k(t)\right)^*\Lambda_{kj}(t)\left(m_k(t) - g_k(t)\right) + Tr\left(\Lambda_{kj}(t)\gamma_k(t)\right).$$

Simulation Studies

The case of one discrete state

We simulated data using description of the FHS. Each individual is characterized by a covariate Y_t, which dynamics is described by the stochastic differential equations (11). For each individual we simulated two-year survival using the quadratic hazardmodel (10). We assumed: i) time independence of a, b, and Q ii) linear age dependence of parameters f and f_1, and iii) Gompetz type function for μ_0. We applied this model to the FHS data and estimated nine parameters. Then we used these parameters for data generation in the simulation study. Using these true parameters we simulated 40 datasets each included 50 year follow-up and longitudinal measurements of Y_t with 82,000 person-years totally. Then, each simulated dataset was used to estimate the model

parameters. The true values for all nine parameters are shown in Table 1 together with characteristics of empirical distributions for each estimated parameter obtained in 40 simulation/estimation studies.

Table 1. The results of the simulation experiment with 40 datasets

	f_a	$f_1 a$	$\mu_0 \times 10^5$	a	f	f_1	θ	$Q \times 10^5$	b
true	0.000	-0.300	1.000	-0.100	80	95	0.100	1.000	5.00
mean	-0.017	-0.295	0.940	-0.098	80.67	93.58	0.099	1.001	4.94
SD	0.087	0.049	0.485	0.014	4.36	10.25	0.017	0.125	0.38
SE	0.014	0.008	0.077	0.002	0.69	1.62	0.003	0.020	0.06

The case of two states

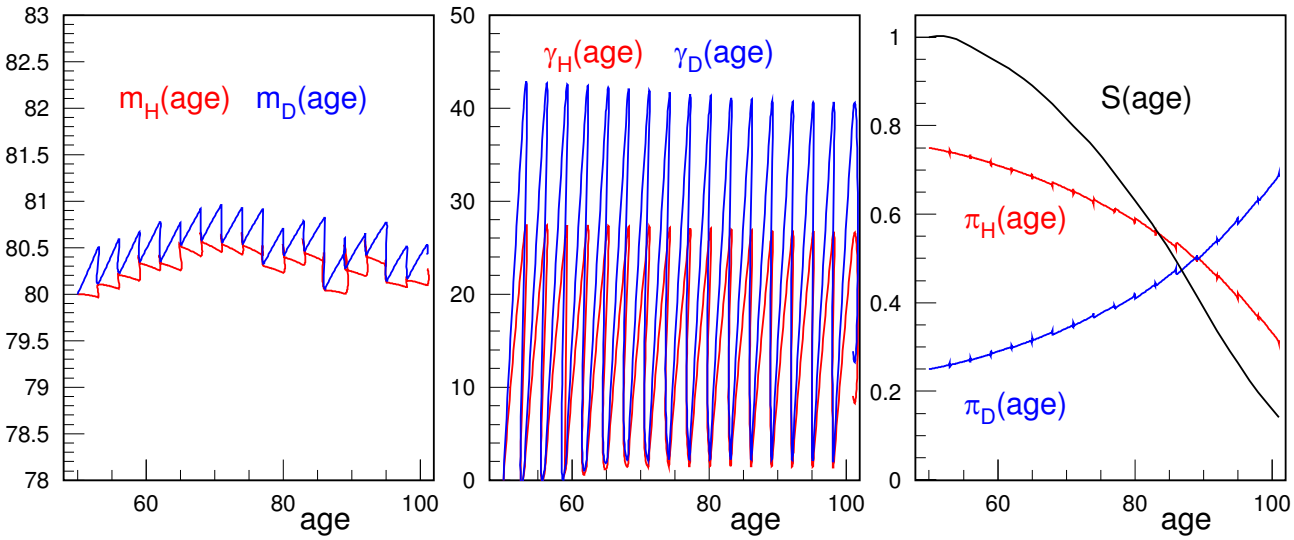
In the second pilot study, we also used the design of the FHS. We assumed that each individual could be characterized by continuously changing physiological indices and be in one of two (healthy or unhealthy) discrete states. Three hazards describing transitions from healthy to unhealthy states ($i = 0$), from healthy state to death ($i = 1$), and from unhealthy state to death ($i = 2$) are modeled as $\mu_j(t,Y) = \mu_j \exp(\theta_i t) + \mu_{iY} \exp(\theta_{iY} t)(Y - f_{iY})^2$.

Table 2. The results of the simulation experiment with 100 datasets (SE=SD/10)

	μ_0	μ_1	μ_2	θ_0	θ_1	θ_2	μ_{0Y}	μ_{1Y}	μ_{2Y}	θ_{0Y}	θ_{1Y}	θ_{2Y}	f_{0Y}	f_{1Y}	f_{2Y}
	10^{-4}	10^{-5}	10^{-5}	10^{-2}	10^{-1}	10^{-1}	10^{-5}	10^{-6}	10^{-6}	10^{-2}	10^{-2}	10^{-2}			
true	1.00	2.00	4.00	5.00	1.00	1.00	1.00	2.00	2.00	5.00	5.00	5.00	80.0	80.0	80.0
mean	1.15	2.07	4.12	5.33	1.00	1.00	1.02	2.24	2.27	4.99	4.96	5.22	80.0	79.9	79.9
SD	1.21	0.81	1.30	1.50	0.01	0.01	0.14	1.07	1.72	0.28	0.92	1.21	0.24	0.84	1.44

Simulating a cohort for observational plan #1

Figure 1 provides basic characteristics of simulated cohorts vs age.



The following set of parameters was used: a_H=-0.05, a_D=-0.03, b_H^2=10, b_D^2=15, f_{1H}=80, f_{1D}=65, μ_{0H}=0.00002, θ_H=0.08, μ_{0D}=0.002, θ_D=0.045, Q_H=0.00001, Q_D=0.00007, Λ_{HD}=0.00005, λ_{0HD}=0.00005, θ_{HD}=0.065, f_H=80, f_D=80, and g_H=72. Then we performed simulation of 10 cohorts using these parameters. For each cohort we reconstructed 8 parameters presented in Table 3. These parameters are responsible for the shape of dynamic trajectories of the covariate in healthe

and disease states and for mortality from these states.

References:

Woodbury MA, Manton KG. (1977) A random walk model of human mortality and aging. *Theor Pop Biol* 11: 37-48.

Yashin AI, Manton KG, Vaupel JW. (1985) Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables. *Theor Pop Biol*. 27: 154-175.

Yashin AI and Manton KG (1997) Effects of unobserved and partially observed covariate processes on system failure: A review of models and estimation strategies. *Statistical Science* 12: 20-34

Yashin AI, Manton KG, Woodbury MA, Stallard E, (1995) The effects of health histories on stochastic process models of aging and mortality. *J. Math. Biol.* 34: 1-16.

Akushevich I, Kulminski A, and Manton K: (2005) Life tables with covariates: Dynamic Model for Nonlinear Analysis of Longitudinal Data. *Mathematical Population Studies*, 12: 51-80.

Yashin A.I., Arbeev K.G., Akushevich I., Kulminski A., Akushevich L., Ukraintseva S.V. (2007) Stochastic model for analysis of longitudinal data on aging and mortality. *Math Biosci* 208: 538-551.

Yashin A.I., Arbeev K.G., Akushevich I., Kulminski A., Akushevich L., Ukraintseva S.V. (2008) Model of hidden heterogeneity in longitudinal data. *Theor Pop Biol.* 73: 1-10.

Arbeev K.G., Akushevich I., Kulminski A.M., Arbeeva L.S., Akushevich L., Ukraintseva S.V., Culminskaya I.V., Yashin A.I. (2009) Genetic model for longitudinal studies of aging, health, and longevity and its potential application to incomplete data. *J Theor Biol* 258: 103-111.