Model of hidden heterogeneity in longitudinal data

Anatoli I. Yashin[1*], Konstantin G. Arbeev[1], Igor Akushevich[1],
Aliaksandr Kulminski[1], Lucy Akushevich[1], Svetlana V. Ukraintseva[1]

[1]Duke University, Center for Population Health and Aging, Durham, NC, USA

[*]Corresponding author: Prof. Anatoli I. Yashin, Duke University, Center for Population Health and Aging, 331 Trent Drive, Room 002, Box 90408, Durham, NC, 27708-0408, USA. Tel. (+1) 919-668-2713, fax: (+1) 919-684-3861. E-mail: aiy@duke.edu

Abstract

Variables measured in longitudinal studies of aging and longevity do not exhaust the list of all factors affecting health and mortality transitions. Unobserved factors generate hidden variability in susceptibility to diseases and death in populations and in age trajectories of longitudinally measured indices. Effects of such heterogeneity can be manifested not only in observed hazard rates but also in average trajectories of measured indices. Although effects of hidden heterogeneity on observed mortality rates are widely discussed, their role in forming age patterns of other aging-related characteristics (average trajectories of physiological state, stress resistance, etc.) is less clear. We propose a model of hidden heterogeneity to analyze its effects in longitudinal data. The approach takes the presence of hidden heterogeneity into account and incorporates several major concepts currently developing in aging research (allostatic load, aging-associated decline in adaptive capacity and stress-resistance, age-dependent physiological norms). Simulation experiments confirm identifiability of model's parameters.

Key words: aging; longevity; quadratic hazard model; heterogeneity; variability; unobserved covariates; longitudinal studies; Framingham Heart Study

1. Introduction

Demographic studies show that hidden differences in susceptibility to death among individuals in a population substantially affect the shapes of the mortality rates at late ages [11,12]. Ignoring the presence of such heterogeneity (e.g., unobserved covariates) results in underestimating regression coefficients in the Cox's proportional hazard model. Hidden variability in other longitudinal characteristics may lead to erroneous conclusions concerning biological regularities of aging-related processes. For example, the average age trajectories of physiological indices may be biased because of effects of mortality selection. For the same reason, the evaluated decline in average resistance to stress can look slower, or become not visible. Hidden heterogeneity may also affect average trajectories of allostatic load, forces of homeostatic regulation, "optimal" trajectories of physiological state, magnitudes of external disturbances, etc.

Many models of hidden heterogeneity in susceptibility to death used in demography and biostatistics are identifiable. It means that distribution of unobserved heterogeneity, regression coefficients and baseline hazard can be evaluated from the data [3]. Models of longitudinal data include description of dynamic properties of basic variables and their connection to mortality. Therefore, they are more complicated than mortality models traditionally used to capture unobserved heterogeneity in the populations (e.g., frailty, random effects, or latent variable models). To our knowledge, there are no results concerning conditions of identifiability in models of heterogeneity for longitudinal data. However, the presence of such heterogeneity is a realistic scenario which cannot be simply ignored in statistical analyses of longitudinal data. Our simulation studies show that parameters of the heterogeneous quadratic hazard model are identifiable in a wide range of respective parameters values.

An important class of models for analyses of longitudinal data uses a biologically-motivated assumption on a quadratic form of the hazard rates. This assumption is also supported by the evidence from epidemiological studies which found the J- or U-shapes of hazards considered as functions of risk factors [15]. These models were developed and intensively used in the studies of longitudinal data [16,17,18,19,5]. The advantageous feature of this approach is that it allows for incorporation of the new findings and ideas appearing in the course of research on aging.

In this paper we introduce the concepts of hidden heterogeneity (discrete frailty) into the quadratic hazard model (QHM). The model (referred to as “QHM with heterogeneity” throughout the paper) allows us to bring together, interrelate, and jointly analyze several fundamental concepts used in different studies of aging-related changes in human organisms. These include the concepts of allostatic load [8], adaptive capacity (homeostenosis) [10,4] and resistance to stresses [9], as well as age-dependent physiological norms [13,1,14]. We performed simulation studies to check the estimation procedure and performance of the model. Application of the model to the Framingham Heart Study (FHS) data on body mass index (BMI) for females illustrates the approach.

2. Model of heterogeneity in longitudinal data

Let Y_t (t is age) be a continuously changing vector of random covariates (e.g., physiological indices) and Z be a hidden heterogeneity variable. It is convenient to describe evolution of Y_t in the form of stochastic differential equation with coefficients depending on Z:

$$dY_t = a(Z,t)\left(Y_t - f_1(Z,t)\right)dt + B(Z,t)dW_t, \quad Y_{t_0}. \qquad (1)$$

Here W_t is a Wiener process independent of the vector of initial conditions Y_{t_0} and the random variable Z. The strength of disturbances W_t is characterized by a matrix of diffusion coefficients $B(Z,t)$. The vector-function $f_1(Z,t)$ (having the same dimension as a vector Y_t) has a meaning of age trajectory of physiological state of the organisms subject to allostasis [7]. The organisms are forced to follow this trajectory by the process of adaptive (homeostatic) regulation. The dependence of this function on Z indicates that mechanisms of allostatic adaptation may differ for groups of individuals characterized by different values of Z. The elements of the matrix $a(Z,t)$ correspond to the rate of adaptive response for any deviation of physiological indices Y_t from $f_1(Z,t)$ for individuals having heterogeneity variable Z.

We illustrate our approach by considering the simplest case, when the random variable Z takes two possible values "0" and "1", $P(Z=1)=p$. The extension to the case with more heterogeneity groups is straightforward. We assume that the conditional distribution of Y_{t_0} given Z is normal with the mean $m\left(k,t_0\right)=m_{k,0}$ and the variance $\gamma\left(k,t_0\right)=\gamma_{k,0}$, $k=0,1$. Let the mortality rate conditional on Y_t and Z be a sum of a baseline (μ_0) and a quadratic hazard:

$$\mu\left(Z,t,Y_t\right)=\mu_0(Z,t)+\left(Y_t-f(Z,t)\right)^*Q(Z,t)\left(Y_t-f(Z,t)\right). \qquad (2)$$

For individuals having the heterogeneity variable Z, the baseline hazard $\mu_0(Z,t)$ characterizes the residual mortality rate, which would remain if the vector of covariates Y_t follows the optimal trajectory (coinciding with the vector-function $f(Z,t)$). The matrix $Q(Z,t)$ is non-negative-definite and symmetric for both values of Z and for all t from the respective interval. The vector-function $f(Z,t)$ is introduced to explicitly characterize age-related changes in the "optimal" physiological state corresponding to the minimum of a hazard rate at a given age and the value of heterogeneity variable Z. It has a meaning of the age-dependent physiological norm for all individuals having the heterogeneity variable Z. It may differ from $f_1(Z,t)$ since the process of allostatic adaptation does not necessarily result in the optimal physiological state.

Such a description corresponds to the assumption that a population under study is a mixture of two subpopulations of individuals (numbered "1" and "0") with initial proportions p and $1-p$, respectively. These subpopulations are characterized by different dynamics of continuously changing covariates and different mortality rates. Let $\tilde{Y}_0^t=Y_{t_0},Y_{t_1},Y_{t_2},\ldots,Y_{t_i}$, $t_i\le t<T$ be a random $i+1$-dimensional vector of observations of the process Y_t at ages $t_0,t_1,t_2,\ldots,t_i$, $t_i\le t<T$. Denote by $\pi(t)=P\left(Z=1\middle|\ \tilde{Y}_0^t,T>t\right)$ the conditional probability that a living individual of age t, having a sequence of measurements $\tilde{Y}_0^t$, belongs to the subpopulation 1. The evolution of $\pi(t)$ starts at age t_0 and first continues at the interval $t_0\le t<t_1$; $t<T$. Using the Bayes formula one can show that $\pi(t)$ satisfies the nonlinear differential equation [17]

$$\frac{d\pi(t)}{dt}=\pi(t)\left(\bar{\bar{\mu}}(t)-\bar{\mu}(1,t)\right), \qquad (3)$$

with the initial condition $\pi(t_0)=p$. Here $\bar{\bar{\mu}}(t)=\pi(t)\bar{\mu}(1,t)+\left(1-\pi(t)\right)\bar{\mu}(0,t)$, and $\bar{\mu}(1,t)$ and $\bar{\mu}(0,t)$ are as follows:

$$\bar{\mu}(k,t)=\mu_0(k,t)+\left(m(k,t)-f(k,t)\right)^*Q(k,t)\left(m(k,t)-f(k,t)\right)+\ Tr\left(Q(k,t)\gamma(k,t)\right), \qquad (4)$$

where $m(k,t)$ and $\gamma(k,t)$, $k=0,1$, are the mean and the variance of the conditional distribution $P\left(Y_t\le y\middle|\ Z=k,T>t\right)$, which satisfy the following ordinary differential equations:

$$\frac{dm(k,t)}{dt}=a(k,t)\left(m(k,t)-f_1(k,t)\right)-2\gamma(k,t)Q(k,t)\ \left(m(k,t)-f(k,t)\right), \qquad (5)$$

$$\frac{d\gamma(k,t)}{dt}=a(k,t)\gamma(k,t)+\gamma(k,t)a(k,t)^*+B(k,t)B(k,t)^*-\ 2\gamma(k,t)Q(k,t)\gamma(k,t), \qquad (6)$$

at the interval $t_0\le t<t_1$, with initial values $m\left(k,t_0\right)=m_{k,0}$; $\gamma\left(k,t_0\right)=\gamma_{k,0},k=0,1$. The

equations (4)-(6) are similar to those derived in [19] in the absence of heterogeneity.

At the age $t = t_1$, $\pi(t)$ jumps because the observation Y_{t_1} brings new information about the value of Z. Using the Bayes rule one can easily calculate the connection between $\pi(t_1)$ and $\pi(t_1-) = \lim_{t\uparrow t_1} \pi(t)$:

$$\pi(t_1) = \frac{\pi(t_1-)\sqrt{\gamma(0,t_1-)}e^{-\frac{\left(Y_{t_1}-m(1,t_1-)\right)^2}{2\gamma(1,t_1-)}}}{\pi(t_1-)\sqrt{\gamma(0,t_1-)}e^{-\frac{\left(Y_{t_1}-m(1,t_1-)\right)^2}{2\gamma(1,t_1-)}} + \left(1-\pi(t_1-)\right)\sqrt{\gamma(1,t_1-)}e^{-\frac{\left(Y_{t_1}-m(0,t_1-)\right)^2}{2\gamma(0,t_1-)}}}. \quad (7)$$

The value $\pi(t_1)$ serves as an initial condition for $\pi(t)$ evolving in accordance with equation (3) at the next age interval: $t_1 \le t < t_2$; $t < T$, and so on. Thus, $\pi(t)$ evolves in accordance with (3) at the age intervals $t_i \le t < t_{i+1}$; $t < T$. The initial values at the beginning of each interval are given by the equation:

$$\pi(t_i) = \frac{\pi(t_i-)\sqrt{\gamma(0,t_i-)}e^{-\frac{\left(Y_{t_i}-m(1,t_i-)\right)^2}{2\gamma(1,t_i-)}}}{\pi(t_i-)\sqrt{\gamma(0,t_i-)}e^{-\frac{\left(Y_{t_i}-m(1,t_i-)\right)^2}{2\gamma(1,t_i-)}} + \left(1-\pi(t_i-)\right)\sqrt{\gamma(1,t_i-)}e^{-\frac{\left(Y_{t_i}-m(0,t_i-)\right)^2}{2\gamma(0,t_i-)}}}. \quad (8)$$

Respectively, $m(k,t)$ and $\gamma(k,t)$, $k = 0,1$, satisfy equations (5) and (6) at the intervals $t_i \le t < t_{i+1}$, with initial values $m\left(k,t_i\right) = Y_{t_i}$; $\gamma\left(k,t_i\right) = 0$, $k = 0,1$. When $t_i \le t = T$,

$$P\left(Z=1\middle|\ \tilde{Y}_0^T, T=t\right) = \pi(T-)\frac{\mu(1,T)}{\bar{\mu}(T)}, \quad \pi(T-) = \lim_{t_n\uparrow t}\pi(t_n)\big|_{t=T}. \quad (9)$$

Thus, if we introduce $\tilde{\pi}(t) = P\left(Z=1\middle|\ \tilde{Y}_0^t, X_0^t\right)$ where $X_t = I(T \le t)$, the trajectory of $\tilde{\pi}(t)$ at the interval $t_i \le t \le T$ can be represented in terms of stochastic differential equation with one jump:

$$d\tilde{\pi}(t) = \tilde{\pi}(t-)\left(\frac{\bar{\mu}(1,t)}{\bar{\bar{\mu}}(t)} - 1\right)\left(dX_t - \bar{\bar{\mu}}(t)dt\right). \quad (10)$$

Here $\bar{\bar{\mu}}(t) = \tilde{\pi}(t)\bar{\mu}(1,t) + \left(1-\tilde{\pi}(t)\right)\bar{\mu}(0,t)$. Note that $\tilde{\pi}(t)I(T>t) = \pi(t)I(T>t)$. The likelihood function of the data is the product of two terms:

$$L_Y = \prod_{j=1}^{N}\prod_{i=0}^{n(j)}\left(\frac{\pi_j(t_i^j-)}{\sqrt{2\pi\gamma_j(1,t_i^j-)}}e^{-\frac{\left(y_j(t_i^j)-m_j(1,t_i^j-)\right)^2}{2\gamma_j(1,t_i^j-)}} + \frac{(1-\pi_j(t_i^j-))}{\sqrt{2\pi\gamma_j(0,t_i^j-)}}e^{-\frac{\left(y_j(t_i^j)-m_j(0,t_i^j-)\right)^2}{2\gamma_j(0,t_i^j-)}}\right), \quad (11)$$

and

$$L_T = \prod_{j=1}^{N}\bar{\bar{\mu}}_j(T_j)^{\delta_j}e^{-\int_0^{T_j}\bar{\bar{\mu}}_j(u)du}, \quad (12)$$

where

$$\bar{\bar{\mu}}_j(t) = \pi_j(t)\bar{\mu}_j(1,t) + \left(1-\pi_j(t)\right)\bar{\mu}_j(0,t), \quad (13)$$

and

$$\bar{\mu}_j(k,t) = \mu_0(k,t) + \left(m_j(k,t) - f(k,t)\right)^* Q(k,t)\left(m_j(k,t) - f(k,t)\right) + Tr\left(Q(k,t)\gamma_j(k,t)\right).$$

Here N is the number of individuals, δ_j is the censoring indicator ($\delta_j = 1$ if j^{th} individual died at age T_j, $\delta_j = 0$ otherwise), $n(j)$ is the number of measurements of the process Y_t for j^{th} individual. The subscript j in $\pi_j(t)$, $y_j(t_i)$, $m_j(k,t)$, $\gamma_j(k,t)$ indicates that respective characteristics refer to j^{th} individual. The symbol $y_j(t_i)$ denotes the value of the random process Y_t measured in individual j at time t_i.

3. Results

3.1. Simulation study for QHM with heterogeneity

We performed a simulation study to check performance of the model in one-dimensional case. In computer simulations, we used a discrete-time version of the model (1)-(13). We assumed that the baseline mortality in (2) is the Gompertz hazard, $\mu_0(k,t) = a_{\mu_0}(k)e^{b_{\mu_0}(k)(t-t_{min})}$, where $t_{min} = 28$, $k = 0,1$. The quadratic hazard terms, $Q(k,t)$, are taken as linear functions of age, $Q(k,t) = a_Q(k) + b_Q(k)\left(t - t_{min}\right)$. To simplify calculations and to reduce the number of parameters, we assumed that the function $f_1(Z,t)$ coincides with the optimal age trajectory of a physiological index, $f(Z,t)$, $f_1(Z,t) = f(Z,t)$. The optimal age-trajectories and the age-related changes in the rate of adaptive regulation are taken as linear functions, $f(k,t) = a_f(k) + b_f(k)(t - t_{min})$ and $-a(k,t) = a_Y(k) - b_Y(k)(t - t_{min})$, where $b_Y(k) > 0$, and the strength coefficient $B(k,t)$ is assumed to be constant, $B(k,t) = \sigma_1(k)$. The initial distribution of Y_{t_0} is normal with the mean $f(k,t_0)$ and the variance $\sigma_0^2(k)$. The initial proportion of individuals in the first subpopulation ($Z = 1$) is denoted by p. Parameters to be estimated in this model are $a_{\mu_0}(k)$, $b_{\mu_0}(k)$, $a_Q(k)$, $b_Q(k)$, $a_f(k)$, $b_f(k)$, $a_Y(k)$, $b_Y(k)$, $\sigma_0(k)$, $\sigma_1(k)$, for $k = 0,1$, and p. Age at entry into the study was simulated as a discrete random variable uniformly distributed over the interval $[28,62]$. The interval between observations of Y_t equals 2 years. The number of observations (surveys) is 25. This structure resembles the Framingham Heart Study (FHS) data [2]. Parameter values were taken close to the estimates of QHM with heterogeneity applied to the FHS data on BMI for females (see section 3.5). We simulated 100 data sets, 2800 individuals in each data set (which is approximately equal to the number of females in the FHS data), and estimated the discrete model for different data sets using the MATLAB's optimization toolbox [6].

The results of this simulation study are shown in Table 1 and Fig. 1. Mean values, standard deviations and minimal and maximal values of the estimated parameters in 100 simulated data sets are presented in Table 1. Estimated trajectories of logarithms of baseline hazard, quadratic hazard terms, optimal age-trajectories of a physiological index, and age-related changes in the adaptive regulation in two subpopulations for 100 simulated data sets with hidden heterogeneity are shown in Fig. 1. Table 1 and Fig. 1 show that the estimation procedure correctly evaluates the parameters of the model and the means of all parameters are close to their true values. Thus, this procedure provides an adequate quality of estimates for a sample size comparable to that of the sex-specific FHS data and allows one to reveal hidden heterogeneity in such data.

Table 1 about here
Fig. 1 about here

3.2. Example: Ignoring hidden heterogeneity induces inconsistency in estimates

To illustrate that ignoring hidden heterogeneity induces inconsistency in parameter estimates, we estimated simulated data sets with hidden heterogeneity (see section 3.1) using the version of QHM without hidden heterogeneity. This model (referred to as "QHM without heterogeneity") is equivalent to equations (1)-(2) without dependence of Y_t and μ on the heterogeneity variable Z:

$$dY_t = a(t)\left(Y_t - f_1(t)\right)dt + B(t)dW_t, \quad Y_{t_0}, \tag{14}$$

$$\mu\left(t, Y_t\right) = \mu_0(t) + \left(Y_t - f(t)\right)^* Q(t)\left(Y_t - f(t)\right), \tag{15}$$

where $\mu_0(t)$, $Q(t)$, $f_1(t)$, $f(t)$, $a(t)$, $B(t)$, and Y_{t_0} are equivalent to respective expressions described in the previous section but without dependence on Z. Details of the likelihood function and the estimation procedure for model (14)-(15) can be found in [20].

The results of this simulation study are shown in Table 1 (section "No Z") and Fig. 2. Table 1 illustrates that QHM without heterogeneity produces parameter estimates that deviate from respective values of parameters in two subpopulations. As a result, the population trajectories of the logarithm of the baseline hazard ($\ln \mu_0(t)$), quadratic hazard terms ($Q(t)$), optimal age-trajectories of a physiological index ($f(t)$), and age-related changes in the adaptive regulation ($a(t)$) deviate from the true trajectories in two subpopulations (Fig. 2). Thus, ignoring hidden heterogeneity leads to incorrect estimates and wrong conclusions.

Fig. 2 about here

3.3. Example: Estimates of hidden heterogeneity in data generated by a model with different structure

Simulation studies described above estimated data sets generated by QHM with heterogeneity. In real applications, however, the underlying model that generates the observed data is usually unknown. Hence, it is important to check whether the model (QHM with heterogeneity) can determine hidden heterogeneity in data generated by a model with different structure. To check this, we generated data using a model with different structure of the process Y_t and mortality μ. Instead of a quadratic hazard, we used a Cox proportional hazard specification for μ:

$$\mu\left(Z, t, Y_t\right) = \mu_0(Z, t) e^{Q(Z,t)|Y_t - f(Z,t)|}. \tag{16}$$

In equation (2), we assumed a quadratic function for $f_1(Z,t)$ instead of a linear one used for QHM with heterogeneity (see section 3.1), $f_1(Z,t) = a_f(Z) + b_f(Z)(t - t_{min}) + c_f(Z)(t - t_{min})^2$. All other functions were taken equal to those described in section 3.1.

We generated 100 data sets using this model (referred to as "Cox model" throughout the text) and estimated these data sets using QHM with heterogeneity described in section 3.1. The results are summarized in Table 2 and Fig. 3. The results indicate that, although QHM with heterogeneity fails to estimate the baseline hazard $\mu_0(Z,t)$ and the function $Q(Z,t)$ because of the difference in the structure of mortality μ in two models, it is able to determine hidden heterogeneity in data and it distinguishes trajectories of optimal values of a physiological index ($f(Z,t)$) and age-related changes in homeostatic capacity ($-a(Z,t)$) in two subpopulations. Initial proportions of individuals in these subpopulations (parameter p) are also estimated correctly.

We also estimated these data using QHM without heterogeneity described in section 3.2. The results are shown in Table 2 (section "No Z") and Fig. 4. In this case, the population trajectories of optimal values of a physiological index ($f(t)$) and age-related changes in the adaptive regulation ($a(t)$) deviate from the true trajectories in two subpopulations (Fig. 4). Thus, the analysis of data generated by the model with different structure using the model with hidden heterogeneity improved the accuracy of calculations and conclusions.

Table 2 about here

Figs. 3-4 about here

3.4. Example: Heterogeneity may mask the decline in stress resistance

Let us assume for simplicity that the optimal physiological state $f(Z,t)$ does not depend on heterogeneity variable and let the conditional mortality rates for the two groups of individuals be $\mu(1,t,Y_t)=\mu_1(t)\left(Y_t-f(t)\right)^2$ and $\mu(0,t,Y_t)=\mu_0(t)\left(Y_t-f(t)\right)^2$. Then the mortality rate conditional on longitudinal data is

$$\bar{\mu}(t,Y_t)=\pi(t)\mu_1(t)\left(Y_t-f(t)\right)^2+(1-\pi(t))\mu_0(t)\left(Y_t-f(t)\right)^2=$$
$$\left(\pi(t)\mu_1(t)+(1-\pi(t))\mu_0(t)\right)\left(Y_t-f(t)\right)^2=\tilde{\mu}(t)\left(Y_t-f(t)\right)^2.$$

Here $\tilde{\mu}(t)=\mu_1(t)\pi(t)+\mu_0(t)(1-\pi(t))$. It is clear that the rate of narrowing of the U-function of risk (which is associated with the slope of $\tilde{\mu}(t)$) may be lower than that in each of the two risk functions.

3.5. Application of QHM with heterogeneity to the FHS data on BMI for females

To illustrate how our approach works in applications to real data, we analyzed data on BMI for females in the FHS data using QHM with heterogeneity. The FHS cohort consists of 5,209 respondents aged 28-62 years residing in Framingham, Massachusetts, between 1948 and 1952 [2]. The FHS cohort is primarily white and has been followed for 50 years for the occurrence of CVD and death through surveillance of hospital admissions, death registries, and other available medical sources. Examination of participants, including an interview, a physical examination, and laboratory tests, has taken place biennially. In our analyses, we used data on BMI (which are available in all 25 exams) for 2,872 females in the FHS.

Fig. 5 illustrates the differences between mortality characteristics in two subpopulations ("$Z=1$" and "$Z=0$"). The first subpopulation ($Z=1$) has a higher baseline mortality at younger ages and a lower baseline mortality at older ages than the second subpopulation ($Z=0$) and two trajectories intersect at age about 55 years. The quadratic hazard term for the first subpopulation is higher and increases faster than that of the second subpopulation. Individuals in two subpopulations have different patterns of "optimal" trajectories of BMI ($f(Z,t)$), i.e., age-specific values of BMI with minimal mortality at respective ages. Individuals from the first subpopulation have an increasing pattern of "optimal" BMI starting from about 23 *kg*/*m*^{2} at age 28, whereas individuals from the second subpopulation have a decreasing pattern of "optimal" BMI starting from about 30 *kg*/*m*^{2} at age 28. Two subpopulations also have different patterns of age-related changes in the adaptive regulation ($-a(Z,t)$). In the first subpopulation, $-a(Z,t)$ declines with age and in the second subpopulation it is almost constant. For comparison, Fig. 5 also shows respective population trajectories estimated by QHM without heterogeneity ("no Z"). It shows, in particular, that the population "optimal" age-trajectory of BMI deviates from the "optimal" trajectories in the subpopulations.

Fig. 6 displays mortality rates ($\mu(Z,t,Y_t)$) and relative risks of death ($RR(Z,t,Y_t)$) over age (t) and values of BMI at age t (Y_t) for two subpopulations. It shows that the mortality rate for the first subpopulation has a more pronounced quadratic pattern (due to higher trajectories of $Q(Z,t)$) and the respective U-shape of mortality risk is narrowing with age (due to the respective increase in $Q(Z,t)$ with age). In the second subpopulation, the mortality rate does not exhibit a visible quadratic pattern. Thus, individuals from this subpopulation are less sensitive to changes in BMI with respect to mortality risks. The narrowing U-shape of risk indicates the aging-related decline in resistance to stress. This finding is in line with the results obtained in animal aging studies, which show a strong connection between the stress resistance and longevity, as well as the decline with age in resistance to many stresses [9]. Both subpopulations revealed similar behavior of the relative risk with increasing age. Contrary to the mortality risk, it decreases with age and manifests a widening U-shape. Such behavior may indicate the increasing role of senescence in the total mortality compared to selected risk factors in aging individuals.

Figs. 5-6 about here

4. Conclusions

The model proposed in this paper puts together different concepts capable of capturing fundamental features of aging-related changes including the notion of allostasis, a decline in adaptive capacity and in resistance to stresses, the aging-related physiological norms, and hidden heterogeneity in longitudinal data, and connects all these concepts to mortality (or incidence) rates. Such a model provides a possibility to develop comprehensive systemic methodology for analyses of available data on aging-related phenomena. Simulation studies showed that the estimation procedure provides an adequate quality of estimates for a moderate sample size comparable to that of the sex-specific FHS data.

Incorporation of a concept of hidden heterogeneity (discrete frailty) allows one to reveal differences in aging-related physiological parameters in individuals represented by distinct heterogeneity groups summarizing unobserved aging- and mortality-related factors (genetic and non-genetic). Individuals may differ in characteristics of allostatic load, forces of adaptive regulation, "optimal" trajectories of physiological state, magnitudes of external disturbances, etc. All such differences may be identified in the proposed model. A statistical analysis of such differences may be performed using the likelihood ratio test comparing the model with equal and non-equal trajectories in two subpopulations.

Ignoring hidden heterogeneity in aging-related characteristics may affect conclusions about regularities of respective processes. Differential selection can produce patterns of mortality or aging-related characteristics for the population as a whole that are qualitatively different from the patterns for respective subpopulations [12]. For example, if the difference between the "optimal" age-trajectories of a physiological index ($f(Z,t)$) in two subpopulations is ignored and the resulting observed age-trajectory in the entire population is taken as a universal “optimal” trajectory, then this “universal trajectory” will be not optimal for individuals from either of the two subpopulations. Therefore, if policy recommendations and health interventions are based on this “universal” trajectory aiming to keep an individual’s age-trajectory close to this “population” trajectory, then this will actually increase the individual’s chances of death due to increasing deviations from the true “optimal” trajectory in a respective subpopulation.

Note that although the conditional distribution of random continuously changing covariates among survivors is not Gaussian, the entire situation can be exactly described in terms of first two conditional moments of two Gaussian distributions and the conditional proportion of individuals in respective groups. The result can be easily extended to include more heterogeneity groups comprising the entire population. Another possible extension of the model is to include a concept of changing frailty. That is, instead of a frailty variable Z (which is fixed for the entire life span of an individual) one can use a hidden jumping process Z_t representing changes in a (discrete) hidden heterogeneity state over time. Such a model can be useful in analyses of longitudinal data where health histories are unobserved or only partly observed. The estimation methods for such types of quadratic hazard models driven by a non-Markovian stochastic process need to be developed and validated in simulation studies for subsequent use in analyses of longitudinal data.

Acknowledgements

The Framingham Heart Study (FHS) is conducted and supported by the NHLBI in collaboration with the FHS Investigators. This manuscript was prepared using a limited access dataset obtained by the NHLBI and does not necessarily reflect the opinions or views of the FHS or the NHLBI. The work on this manuscript was supported by the NIH/NIA grants 1R01 AG028259-01, 1RO1-AG-027019-01 and 5PO1-AG-008761-16.

References

[1] Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L.Jr, Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T.Jr, Roccella, E.J., 2003. Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. National Heart, Lung, and Blood Institute; National High Blood Pressure Education Program Coordinating Committee. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension 42(6), 1206–1252.

[2] Dawber, T.R., Kannel, W.B., 1958. An epidemiologic study of heart disease: the Framingham Study. Nutr. Rev. 16, 1–4.

[3] Elbers, C., Ridder, G., 1982. True and spurious duration dependence: the identifiability of the proportional hazards model. Review of Economic Studies 49, 403–409.

[4] Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S.K., Johnson, T.E., 2002. Transcriptional profile of aging in C. elegans. Curr. Biol. 12(18), 1566–1573.

[5] Manton, K.G., Yashin, A.I., 2000. Mechanisms of Aging and Mortality: A Search

for New Paradigms. Odense Monograph on Population Aging No. 7. Odense University Press, Odense, Denmark.

[6] Math Works Inc. (Eds.), 2004. Optimization Toolbox for Use with MATLAB. User's guide. Version 3. The Math Works, Inc., Natick, MA.

[7] McEwen, B.S., Wingfield, J.C., 2003. The concept of allostasis in biology and biomedicine. Horm. Behav. 43, 2–15.

[8] Seeman, T.E., McEwen, B.S., Rowe, J.W., Singer, B.H., 2001. Allostatic load as a marker of cumulative biological risk: MacArthur Studies of Successful Aging. Proc. Natl. Acad. Sci. USA. 98, 4770–4775.

[9] Semenchenko, G.V., Khazaeli, A.A., Curtsinger, J.W., Yashin A.I., 2004. Stress resistance declines with age: analysis of data from a survival experiment with Drosophila melanogaster. Biogerontology 5, 17–30.

[10] Troncale, J.A., 1996. The aging process. Physiologic changes and pharmacologic implications. Postgrad. Med. 99(5), 111–114, 120–122.

[11] Vaupel, J.W., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 9, 439–454.

[12] Vaupel, J.W., Yashin, A.I., 1985. Heterogeneity's ruses: some surprising effects of selection on population dynamics. Am. Stat. 39, 176–185.

[13] Westin, S., Heath, I., 2005. Thresholds for normal blood pressure and serum cholesterol. BMJ 330(7506), 1461–1462.

[14] WHO, 2000. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. World Health Organ. Tech. Rep. Ser. 894, I–XII, 1–253.

[15] Witteman, J.C.M., Grobbee, D.E., Valkenburg, H.A., Hemert, A.M. van, Stijnen, Th., Burger, H., Hofman, A., 1994. J-shaped relation between change in diastolic blood pressure and aortic atherosclerosis. Lancet 343, 504–507.

[16] Woodbury, M.A., Manton, K.G., 1977. A random-walk model of human mortality and aging. Theor. Popul. Biol. 11, 37–48.

[17] Yashin, A.I., 1985. Dynamics in survival analysis: conditional Gaussian property vs. Cameron-Martin formula. In: Krylov, N.V., Lipster, R.Sh., Novikov, A.A. (Eds.), Statistics and control of stochastic processes, Springer, New York, pp. 446–475.

[18] Yashin, A.I., Manton, K.G., Vaupel, J.W., 1985. Mortality and aging in heterogeneous populations: a stochastic process model with observed and unobserved variables. Theor. Popul. Biol. 27, 159–175.

[19] Yashin, A.I., Manton, K.G., 1997. Effects of Unobserved and Partially Observed Covariate Processes on System Failure: A Review of Models and Estimation Strategies, Statistical Science 12(1), 20–34.

[20] 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., in press. DOI: 10.1016/j.mbs.2006.11.006.

Figure legends:

Fig. 1: Estimated trajectories of logarithms of baseline hazard ($\ln \mu_0(Z,t)$), quadratic hazard terms ($Q(Z,t)$), optimal values of a physiological index ($f(Z,t)$) and age-related changes in homeostatic capacity ($-a(Z,t)$) in two hypothetical subpopulations ("$Z=1$, est." and "$Z=0$, est.") for 100 simulated data sets with hidden heterogeneity evaluated by QHM with heterogeneity. Respective true trajectories in two subpopulations are denoted as "$Z=1$, true" and "$Z=0$, true".

Fig. 2: Estimated population trajectories of logarithm of baseline hazard ($\ln \mu_0(t)$), quadratic hazard term ($Q(t)$), optimal values of a physiological index ($f(t)$) and age-related changes in homeostatic capacity ($-a(t)$) for QHM without heterogeneity ("no Z, est.") evaluating 100 data sets simulated by QHM with heterogeneity. Respective true trajectories in two hypothetical subpopulations for QHM with heterogeneity are denoted as "$Z=1$, true" and "$Z=0$, true".

Fig. 3: Estimated trajectories of logarithms of baseline hazard ($\ln \mu_0(Z,t)$), quadratic hazard terms ($Q(Z,t)$), optimal values of a physiological index ($f(Z,t)$) and age-related changes in homeostatic capacity ($-a(Z,t)$) in two hypothetical subpopulations for QHM with heterogeneity ("$Z=1$, QHM" and "$Z=0$, QHM") evaluating 100 data sets simulated by Cox model. Respective true trajectories for Cox model are denoted as "$Z=1$, Cox" and "$Z=0$, Cox".

Fig. 4: Estimated population trajectories of logarithm of baseline hazard ($\ln \mu_0(t)$), quadratic hazard term ($Q(t)$), optimal values of a physiological index ($f(t)$) and age-related changes in homeostatic capacity ($-a(t)$) for QHM without heterogeneity ("no Z, QHM") evaluating 100 data sets with hidden heterogeneity simulated by Cox model. Respective true trajectories for Cox model are denoted as "$Z=1$, Cox" and "$Z=0$, Cox".

Fig. 5: Estimated trajectories of logarithms of baseline hazard ($\ln \mu_0(Z,t)$), quadratic hazard terms ($Q(Z,t)$), optimal values of a physiological index ($f(Z,t)$) and age-related changes in homeostatic capacity ($-a(Z,t)$) in two subpopulations ("$Z=1$" and "$Z=0$") for QHM with heterogeneity applied to the FHS data on BMI for females. Population estimates of respective characteristics for QHM model without heterogeneity ("no Z") are shown for comparison.

Fig. 6: Estimated mortality rates ($\mu(Z,t,Y_t)$) and relative risks of death ($RR(Z,t,Y_t)$, logarithmic scale) over age (t) and values of a physiological index at age t (Y_t) in two subpopulations ("$Z=1$" and "$Z=0$") for QHM with heterogeneity applied to the FHS data on BMI for females. Thick black lines denote optimal age-trajectories of BMI ($f(Z,t)$) in respective subpopulations.

Tables:

Table 1. Means, standard deviations (STD), minimal (MIN) and maximal (MAX) values of parameter estimates in 100 simulated data sets for QHM with heterogeneity. Values for QHM without heterogeneity ("No Z") are given for comparison.

$Z=1$:	$a_{\mu_0}\cdot 10^4$	b_{μ_0}	$a_Q\cdot 10^4$	$b_Q\cdot 10^4$	a_Y	$b_Y\cdot 10^4$	σ_0	σ_1	a_f	b_f	p
Mean	2.41	0.090	0.29	0.077	0.049	5.81	3.00	0.90	22.99	0.080	0.70
STD	0.48	0.004	0.36	0.014	0.002	0.52	0.05	0.004	0.11	0.005	0.003
MIN	1.49	0.083	0.00	0.025	0.043	4.12	2.87	0.89	22.75	0.068	0.69
MAX	3.75	0.098	1.77	0.102	0.054	6.81	3.13	0.91	23.24	0.092	0.71
True:	**2.4**	**0.09**	**0.2**	**0.08**	**0.05**	**6.0**	**3.0**	**0.9**	**23.0**	**0.08**	**0.7**

$Z=0$:	$a_{\mu_0}\cdot 10^4$	b_{μ_0}	$a_Q\cdot 10^4$	$b_Q\cdot 10^4$	a_Y	$b_Y\cdot 10^4$	σ_0	σ_1	a_f	b_f
Mean	0.96	0.121	0.14	0.009	0.051	0.75	5.39	1.60	29.98	-0.029
STD	0.29	0.006	0.16	0.007	0.003	0.87	0.14	0.01	0.31	0.010
MIN	0.48	0.110	0.00	-0.006	0.046	0.00	5.09	1.57	29.16	-0.057
MAX	1.76	0.134	0.58	0.024	0.064	3.57	5.86	1.63	30.82	-0.006
True:	**1.0**	**0.12**	**0.1**	**0.01**	**0.05**	**0.5**	**5.4**	**1.6**	**30.0**	**-0.03**

No Z:	$a_{\mu_0}\cdot 10^4$	b_{μ_0}	$a_Q\cdot 10^4$	$b_Q\cdot 10^4$	a_Y	$b_Y\cdot 10^4$	σ_0	σ_1	a_f	b_f
Mean	2.79	0.092	0.001	0.032	0.035	1.07	4.56	1.15	25.24	0.040
STD	0.41	0.003	0.007	0.004	0.002	0.58	0.07	0.005	0.13	0.005
MIN	1.87	0.086	0.00	0.017	0.031	0.00	4.38	1.14	24.95	0.030
MAX	4.05	0.099	0.07	0.043	0.041	2.73	4.80	1.16	25.53	0.049

Table 2. Means, standard deviations (STD), minimal (MIN) and maximal (MAX) values of parameter estimates for QHM with heterogeneity evaluating 100 data sets simulated by Cox model. Values for QHM without heterogeneity ("No Z") estimating these data are given for comparison.

$Z=1$:	$a_{\mu_0}\cdot 10^4$	b_{μ_0}	a_Q	$b_Q\cdot 10^2$	a_Y	$b_Y\cdot 10^4$	σ_0	σ_1	a_f	b_f	p
Mean	1.16	0.111	0.00	0.001	0.050	5.90	3.00	0.90	23.10	0.064	0.70
STD	0.20	0.003	0.00	0.0001	0.003	0.61	0.05	0.004	0.10	0.004	0.004
MIN	0.78	0.102	0.00	0.0008	0.041	3.41	2.86	0.89	22.88	0.055	0.69
MAX	1.87	0.118	0.00	0.001	0.056	7.06	3.16	0.91	23.31	0.074	0.71
True:	**2.4**	**0.09**	**0.03**	**0.3**	**0.05**	**6.0**	**3.0**	**0.9**	**23.0**	**0.08**	**0.7**

$Z=0$:	$a_{\mu_0}\cdot 10^4$	b_{μ_0}	a_Q	$b_Q\cdot 10^2$	a_Y	$b_Y\cdot 10^4$	σ_0	σ_1	a_f	b_f
Mean	0.49	0.150	0.00	0.001	0.052	1.27	5.39	1.60	30.02	-0.024
STD	0.14	0.006	0.00	0.0001	0.004	1.31	0.14	0.01	0.29	0.010
MIN	0.22	0.135	0.00	0.001	0.044	0.00	5.03	1.58	29.15	-0.048
MAX	0.93	0.166	0.00	0.001	0.064	4.40	5.67	1.63	30.69	0.006
True:	**1.0**	**0.12**	**0.04**	**0.4**	**0.05**	**0.5**	**5.4**	**1.6**	**30.0**	**-0.03**

No Z:	$a_{\mu_0}\cdot 10^4$	b_{μ_0}	a_Q	$b_Q\cdot 10^2$	a_Y	$b_Y\cdot 10^4$	σ_0	σ_1	a_f	b_f
Mean	1.80	0.107	0.00	0.001	0.036	2.20	4.62	1.13	25.39	0.025
STD	0.22	0.002	0.00	0.0001	0.003	0.74	0.06	0.01	0.12	0.004
MIN	1.33	0.102	0.00	0.001	0.030	0.12	4.46	1.12	25.04	0.015
MAX	2.38	0.113	0.00	0.001	0.044	4.40	4.79	1.15	25.69	0.037

Figures:

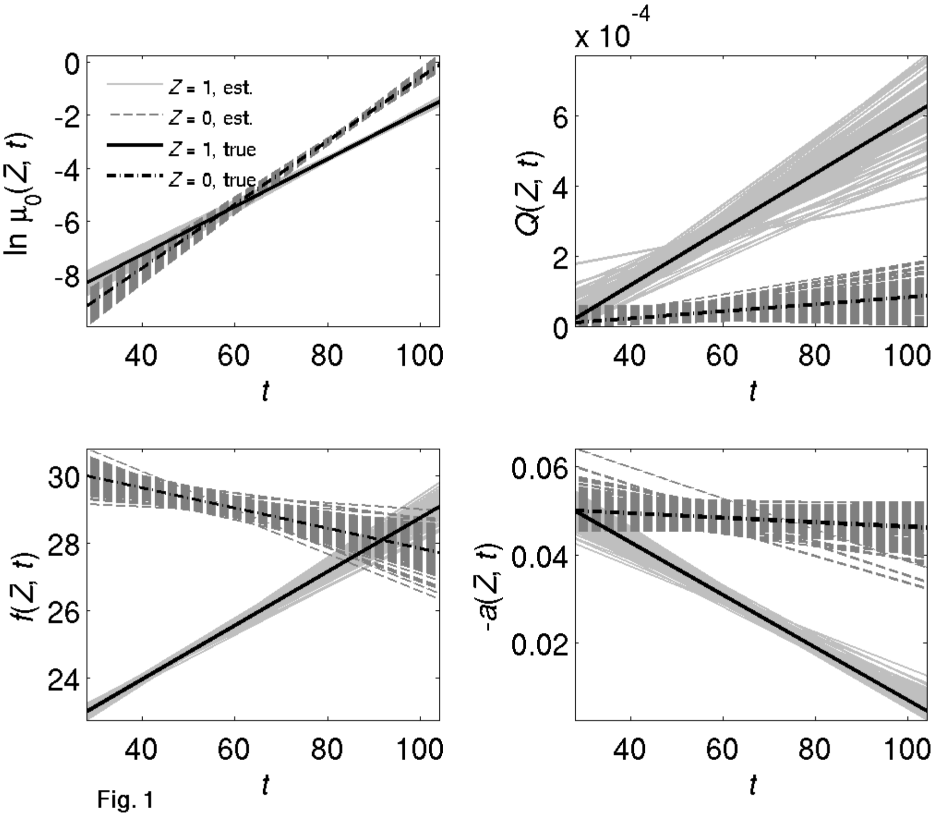

Fig. 1

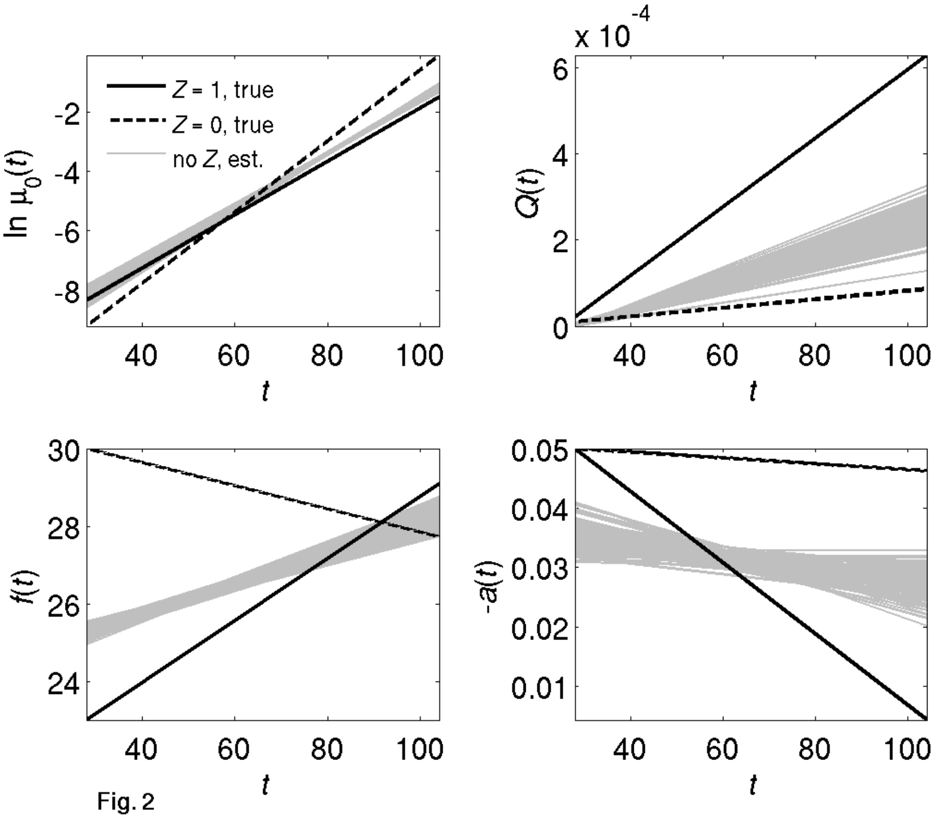

Fig. 2

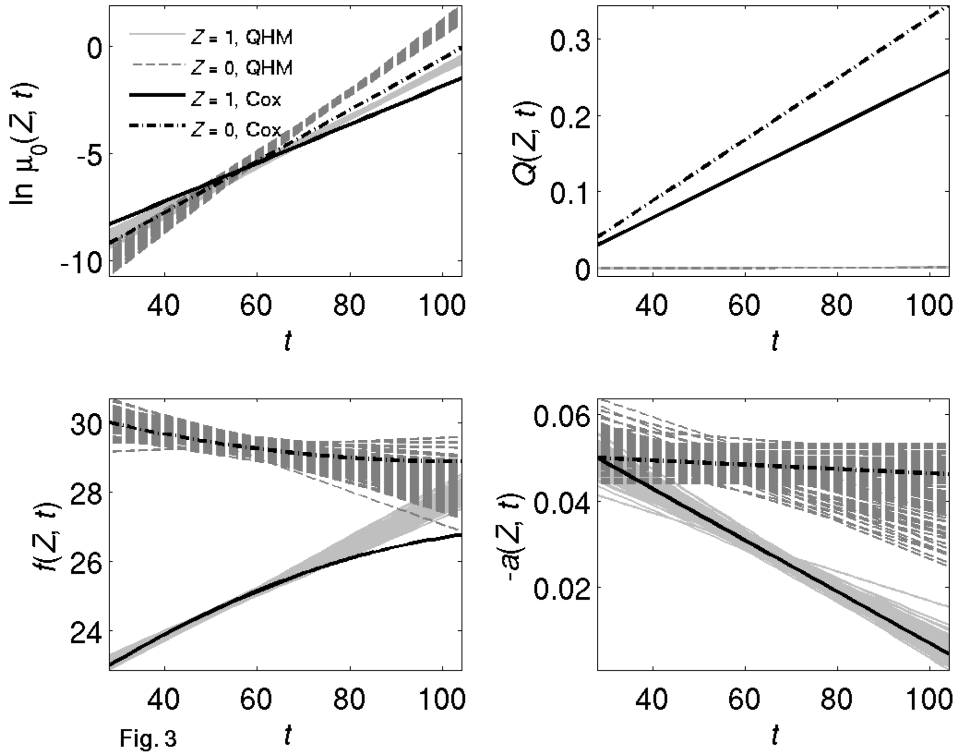
Z = 1, QHM
Z = 0, QHM
Z = 1, Cox
Z = 0, Cox
ln μ0(Z, t)
0
-5
-10
Q(Z, t)
0.3
0.2
0.1
0
f(Z, t)
30
28
26
24
-a(Z, t)
0.06
0.04
0.02
40
60
80
100
t

Fig. 3

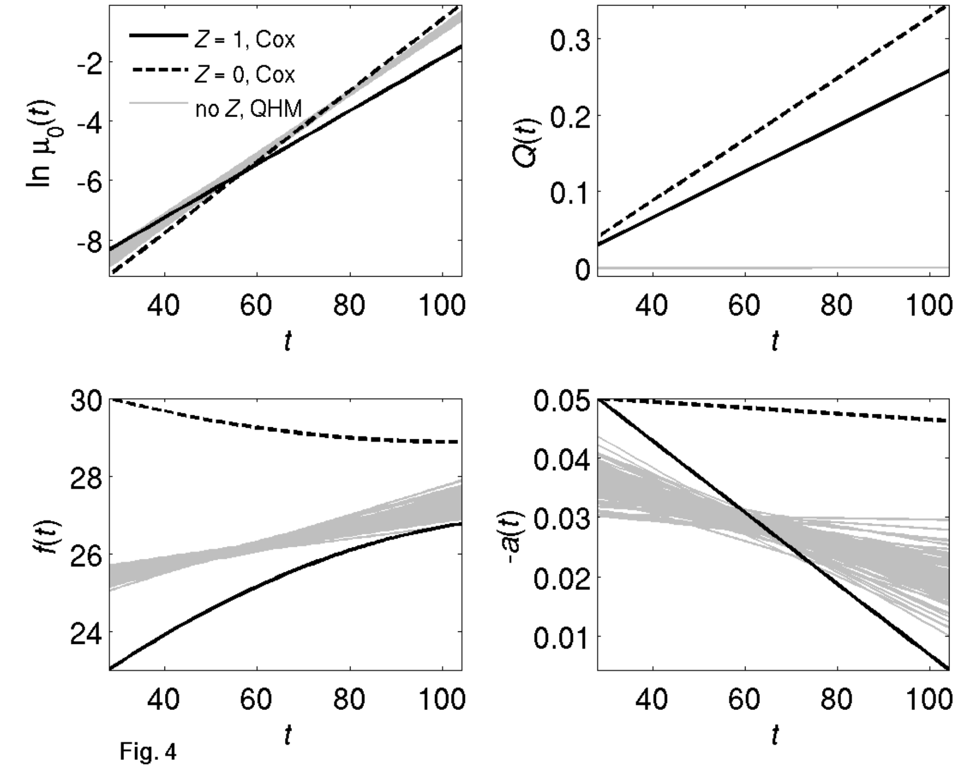
Z = 1, Cox
Z = 0, Cox
no Z, QHM
ln μ0(t)
-2
-4
-6
-8
Q(t)
0.3
0.2
0.1
0
f(t)
30
28
26
24
-a(t)
0.05
0.04
0.03
0.02
0.01
40
60
80
100
t

Fig. 4

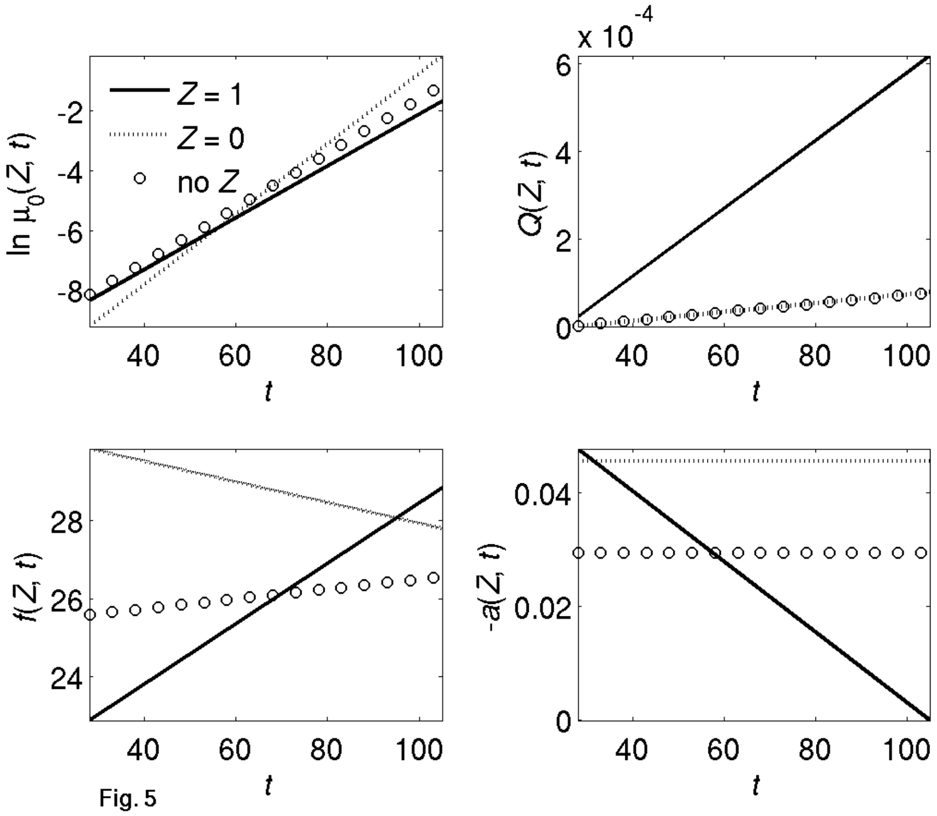
Z = 1
Z = 0
no Z
ln μ0(Z, t)
-2
-4
-6
-8
x 10^-4
Q(Z, t)
6
4
2
0
f(Z, t)
28
26
24
-a(Z, t)
0.04
0.02
0
40
60
80
100
t

Fig. 5

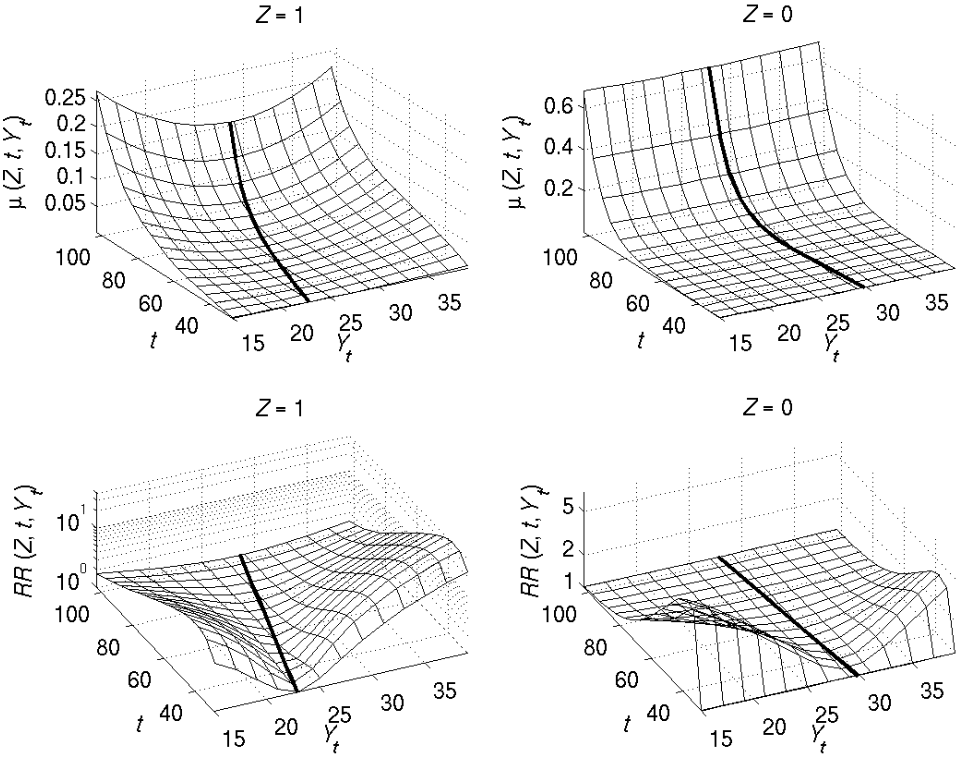
Z = 1
Z = 0
μ (Z, t, Y_t)
0.25
0.2
0.15
0.1
0.05
0.6
0.4
0.2
RR (Z, t, Y_t)
10^1
10^0
5
2
1
100
80
60
40
t
15
20
25
30
35
Y_t

Fig. 6