MODELING DISEASE DYNAMICS AND SURVIVOR FUNCTIONS BY SANOGENESIS CURVES

Bart A.G.,* Bart V.A.,† Steland A.,‡ Zaslavskiy M.L.§

Abstract

We propose to analyse the development of a disease by sanogenesis curves which model the result of interacting exitatory and inhibitory factors on a disease. Assuming that high-dimensional data describing the disease course are driven by a latent complex-valued Gaussian process with Markovian structure, we can identify the sanogenesis curve as the real part of the covariance function of the latent process. By applying techniques of stochastic process theory and partially inverse functions theory this finding allows to estimate the model parameters. In addition, the sanogensis curve also suggests a new model for survival times, where failures (deads) are only observed during critical time periods (crises) defined by the sanogenesis curve. We illustrate our approach by analyzing two real data sets from medicine.

Keywords: Sanogenesis curve, survivor function, partially inverse function, factor analysis, correlation function, critical and latent time points, confidence regions

*St.-Petersburg State University.
†Research Institute of Cardiology, St.-Petersburg.
‡Ruhr-Universität Bochum.
§St.-Petersburg State University.

1 Introduction

This paper is devoted to a new approach to analyze the development of a disease and related survivor functions. Our research is primarily motivated by medical applications where often high-dimensional data about the development of a disease is given and should be used in the analysis. Given a generic survival time T, i.e., a non-negative random variable (r.v.), the most important statistical quantities are the survivor function defined by $H(t) = P(T > t)$, $t \in \mathbb{R}$ and the hazard function $h(t) = -d \ln H(t)/dt$. There exists a rich literature discussing approaches to model these quantities. We refer to Barlow and Proschan (1975), Cox and Oakes (1984), Gavrilov and Gavrilova (1991), Klein and Goel (1992), and the references given there. These classic approaches share the following features, which may be unpleasant for certain applications.

First, most classic approaches assume smooth and rather simple survivor functions. Medical data of disease processes are often characterized by a discrete-continuous character reflecting the dynamics of decay periods and latent periods. In particular, these latent periods play an essential role in the dynamics of many diseases sometimes determining their course, see Bart, Bondarenko, and Boiko (1980). One may consider non-parametric models, e.g., the Kaplan-Meier estimates of the survivor function, cf. Cox and Oakes (1984) and Glantz (1994), but in many cases an underlying parametric model taking account of discontinuities and latent periods is more appealing.

Second, survivor functions describe only the result of a complex regulatory interaction process of the involved (biological) system. In the sequel we call these underlying processes basic processes. The general problem, however, has an integral character. Our aim is to provide a relatively detailed framework for the system behavior which drives the distribution of the observed survival times and additional variables which describe the disease dynamics. In real applications the survivor functions are usually accompanied by rich information about the basic processes. In that case, the medical data usually consists of high-dimensional basic information about the patient's disease development, e.g., blood measurements, heart rate, CD4 cell counts playing an important role in the immune system, etc. Consequently, we have a large number of additional observed variables for each patient. Traditionally, that kind of information is considered as additional data and is taken into account in the form of explanatory variables as in the proportional hazard model, see Cox and Oakes (1984). In that model the shape of the hazard function remains the same, but the initial function $h_0(t)$ is multiplied by a factor, $\psi(\beta, \mathbf{z})$, depending on a parameter vector β (so that $\psi(0, \mathbf{z}) = 1$) and on a vector $\mathbf{z}$ of explanatory variables (covariats), which alone contains all the additional information, yielding the model $h(t) = \psi(\beta, \mathbf{z})h_0(t)$.

The present paper proposes a completely different approach to deal with these problems. Following Bart, Bondarenko, and Boiko (1980), Bart and Alekseyeff (1999), and Bart (2002) a parametric model of the survivor function is formulated, which is based on an analysis of the dynamics due to the main regulatory interactions in the disease course. In a first stage we propose to perform a factor analysis to reduce the complexity of the variables describing the disease course to two main factors. Factor analyses are quite common in medicine, see e.g. Lloyd and Ledermann (1984). The resuling two-dimensional process is used to estimate a parametric model for the sanogensis curve. We regard that function as the main subject of an analysis of a disease, and as an important starting point to derive survivor functions. In medical-biological systems the sanogenetic (inhibitory) factors retard a pathological process driven by exitatory factors. When confronted with a disease, an organism's immune systems initiates regulatory mechanisms. The sanogensis functions describes the dynamics of these interactions and has therefore a medical interpretation.

Our approach is to derive a survivor function directly from an (estimated) sanogenesis curve by a transformation which essentially puts an upper bound on those oscillation of the sanogenesis curve which are under control of the organism. Mathematically, the transformation to a survivor function is given by a special inversion operator applied to the sanogensis curve. That operator is parameterized by a control parameter corresponding to all pre-images of a function with discontinuities. This parameter plays an important role in our analysis. In particular, in the survivor function model under consideration it permits to take allowance for different variants of a disease course and also to recognize the critical moments in the disease dynamics. That important aspect is new compared to classic approaches.

Our approach gives a survivor function model using the maximal information available from the data describing the disease course. Its concordance with the specific survivor function is a problem which has to be solved in a collaboration of the statistician with the practitioner not only by means of statistics but also with a high portion of semantics. E.g., for the surgery data set discussed below in detail we were able to show that our survivor function is in agreement with the Kaplan-Meier estimate by means of a statistical homogeneity test.

The paper is organized as follows. The model and the application of the partial inversion operator are presented in Section 2. An explicit method to construct a partially inverse function which serves our needs is briefly described in appendix. The general theory of the functions partial inversion and its different applications one can be found in Bart (2002), Bart, Klochkova, and Kozhanov (1993), and Bart, Alekseyeff, and Bochkina (2000). In Section 3 the pointwise estimates of the parameters of the model are presented and confidence intervals for them are constructed. We illustrate our method in Section 4 by a simulated data example and a real data set. Proofs of the main results are given in the appendix.

2 Sanogenesis curves and survivor functions

In this section we carefully explain our model and its components. Denote by $(T, \{Z_1, \ldots, Z_k\})$ the generic observations of a patient, where T denotes the survival time and $Z_1, \ldots, Z_k$ is a time series of p-dimensional random vectors (r.v.s) describing the development of a disease in terms of important clinical variables. We assume that the p-dimensional r.v.s Z_t satisfies a two-factor model, where the two independent latent factors form a complex-valued Markovian normal stationary process (CMNS process). The real part of its covariance function will be identified with a sanogenesis curve denoted by $\mathcal{S}(t)$, which can be derived by modeling the regulatory mechanism of pathogenetic factors and (organic) factors inhibiting the disease by a certain deterministic system of differential equations. By applying an appropriate double inversion operator to the sanogenesis function we are also in a position to construct a survivor function $H(t)$ providing a distributional model for T, i.e., $H(t) = P(T > t), t \in \mathbb{R}$. Under that survivor function one can die with possitive probability during certain intervals (crises) which occur periodically and are defined by the sanogensis curve.

We start with a derivation of sanogenesis curves from basic principles describing the interaction of pathogenetic and sanogenetic factors of a disease. The inversion operator which transforms a sanogenesis curve to survivor function is described in Subsection 2.2. The next step is to formulate the two-factor model relating the observations Z_t to a latent Gaussian complex-valued process, whose covariance function comprises the sanogenesis function.

2.1 Sanogenesis curves

The sanogenesis curve summarizes the effects of organic regulatory mechanisms on the disease course. We assume that it is given by

$$\mathcal{S}(t) = \sigma^2 e^{-\eta t} \cos \tau t, \quad t > 0, \quad \eta > 0, \tag{1}$$

where η, τ are the parameters which determine a process course character, including its critical periods. This function was introduced for survivor function analysis in Bart, Bondarenko, and Boiko (1980) and was called a sanogenesis (compensation) curve.

The explicit form of the sanogenesis curve (1) can be justified from different assumptions. In Bart, Bondarenko, and Boiko (1980) the function $\mathcal{S}(t)$ (more exactly its first nonzero branch) was derived as a survivor function model of patients with glomerulonephritis of the chronic nephritic insufficiency stage by considering the disease in the context of the regulatory system organ - organism.

The aim of such an analysis is to take account of interactions of *pathogenetic* and *sanogenetic* factors both at the organ level, denoted by $((p(t), s(t))^T = x_t)$, and of the organism, denoted by $((\mathcal{P}(t), \mathcal{S}(t))^T = \Pi(t) = \Pi)$.

Using the principle of least action from mechanics, we obtain the following system of differential equations with respect to Π.

$$\dot{\Pi} = \mu \Pi, \quad \text{where } \mu = \begin{bmatrix} \eta & \tau \\ -\tau & \eta \end{bmatrix}.$$

The homeastaz conditions for the patho- and sanogenetic factor interaction at the organism level, $(\mathcal{P}(t) + \mathcal{S}(t) = 0)$, we obtain the explicit representation

$$\mathcal{S}(t) = e^{-\eta(t-T_0)} \cos \tau(t - T_0), \qquad T_0 \le t \le T_L,$$

for the sanogenesis curve, where T_0 is the end of latent stage and T_L the first (terminal) time moment where $\mathcal{S}(T_L) = 0$.

Critical time points, denoted by T_k, are defined as solutions of the equation

$$e^{-\eta(T_k - T_0)} = \cos\tau(T_k - T_0) = \theta, \tag{2}$$

where θ denotes the common value for the right and left parts in (2). The parameter $\theta = \theta(\eta, \tau)$ reflects the speed of disease progress.

2.2 Construction of a survivor function from a sanogenesis curve

The sanogenesis curve allows to define critical time points and time intervals (crises) where the pathogenetic factors start to dominate the sanogenetic factors, and that pattern is periodically observed during the development of the disease. A plausible model for many diseases is to assume that one can die with positive probability only during these periodic time intervals where the exitatory factors start to dominate the inhibitory factors. This suggests to use the sanogenesis curve to define a survivor function.

Obviously, only in special cases the function $\mathcal{S}(t)$ defines a survivor function over the time period of interest. The general case when the function $\mathcal{S}(t)$ fails to be decreasing and shows a periodic pattern is of more interest. As an example we discuss in Section 4.3 a real data set of patients with arterial hypertension, where the disease process has several stages with corresponding crises where the probability to die is high.

Consequently, the sanogenesis curve cannot be used as a survivor function directly. We propose to map $\mathcal{S}(t)$, which represents a rhythm structure, to an element $H(t)$ of the set of survivor functions via a transformation which is given by $H = G_\gamma \circ G_{\gamma'}(\mathcal{S})$ where G_γ is the partial inversion operator with partiality parameter $\gamma \in (0, 1]$ described in detail in the appendix. We also write $\mathcal{S}^{-2}_{\gamma,\gamma'}$ for the double invers $G_\gamma \circ G_{\gamma'}(\mathcal{S})$ of $\mathcal{S}$. Of primary interest are the cases

$$H_\gamma(t) = \mathcal{S}^{-2}_{\gamma,1}(t); \quad H_0(t) = \mathcal{S}^{-2}_{0,0}(t); \quad H_1(t) = \mathcal{S}^{-2}_{1,1}(t).$$

Here $H_0(t)$ corresponds to the double left inversion of $\mathcal{S}(t)$ and $H_1(t)$ to the double right inversion. The first inversion involves the partiality parameter γ. If $0 < \gamma \leq 1$ then we always consider the second inversion to be the right one.

Figure 1 illustrates that inversion process of the sanogenesis function $\mathcal{S}(t)$. Shown are the right inversion $H_1(t)$, the left $H_0(t)$, an inversion $H_\gamma(t)$ using the parameter $\gamma = 0, 5$.

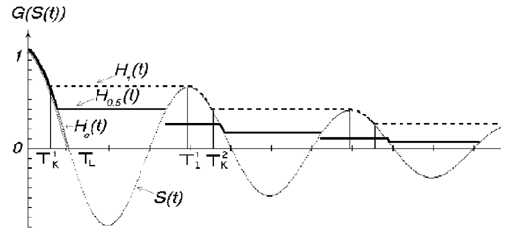

Figure 1: Functions $S(t)$, $H_1(t)$, $H_0(t)$ and $H_{0.5}(t)$.

In our survivor function model the decay stages, where the function decreases, and the latent stages, where the function remains constant, alternate. The critical time points T_k^j and the latent ones T_l^j are defined as the boundary between the latent and decay stages. The time point T_L is a terminal one in the most unfavorable course. The properties of these critical time points, methods of calculating and testing of the model by some real systems are presented in Bart, Alekseyeff and Bochkina (2000) and Bart (1999, 2002). It is easy to verify that the distance between neighboring critical moments of time $\{T_k^{j+1} - T_k^j,\, j = 1, 2 \ldots\}$ forms an arithmetic progression and the jumps between the meanings of the function in them, namely a geometric one. The same applies to the latent points.

Remark 1 *When applying the model the interpretation of the time structure is of some importance. Usually, the symptoms of a disease appear asynchronously and the associated event times are considerably disturbed by random noise. Thus, one should primarily consider the private time of the disease and should base conclusions upon the most contrasting symptoms which fix the stages of the disease process, as modeled by the sanogenesis curve.*

For example, the stages of arterial hypertension (AH) correspond to the quite new qualitative states connected with some functional changes in vessels and target-organ (AH-1), and then with some organic ones (AH-2). The disease stages evolution sequence corresponds a transition from the boundary AH

(BAH) to the next stages of AH. In the section 4.2 four stages are analyzed: the norm, BAH, AH-1 and AH-2. The time moments determining these stages may be considered as the critical points but their concordance with the common (laboratory) time is not simple. The wound processes in the section 4.3 have a rhythm which is better agreed with the lab-time. There a transition from one qualitative stage to another is estimated at about a week, so the initial moments of these transitions are taken as latent points. In general, the coordination of an internal time of disease course and a lab-time is a delicate problem.

2.3 Two-factor model

In general, the data available for each patient is high-dimensional and the variables are highly correlated. Therefore we propose to reduce the dimensionality by a factor model where the correlation structure is explained by a small set of latent factors. These factors describe the development of the disease. We assume that the generic data of a patient, $\{Z_t\}$, where Z_t is a p-dimensional r.v. with mean μ_t, satisfies a two-dimensional factor model at each time point t,

$$Z_t - \mu_t = \mathbf{\Lambda}_t F_t + e_t, \qquad t = 1, \ldots, k,$$

where Λ_t is a $p \times 2$ matrix of regression weights (loadings), and $F_t = (U_t, V_t)'$ is a vector of unobserved independent latent factors, and e_t is an error term. It is assumed that $EF_t = 0$ and $Ee_t = 0$.

Remark 2 *Explaining the correlation structure by only two independent sources seems to be restrictive, but the data analyses discussed below in greater detail suggest that the evolution of many diseases can be nicely captured in many cases. In our data analyses it turned out to be very useful to consider also multi-factor models and select those two factors which provide a meaningful interpretation.*

Remark 3 *Alternatively, one apply classic principle component analysis (PCA) or independent component analysis (ICA), cf. Hyvarinen, Karhunen, and Oja (2001) and Chen and Bickel (2003), to find appropriate latent factors.*

2.4 CMNS process for the latent factors

The two-factor model above defines a complex-valued process in discrete time,

$$X_t = U_t + iV_t, \qquad t = 1, \ldots, k,$$

where $i^2 = -1$. We assume that $\{X_t\}$ is a Markovian Gaussian process, i.e., particularly we have

$$P(X_t \in A | X_{t-1}, X_{t-2}, \ldots) = P(X_t \in A | X_{t-1})$$

for any measurable set $A \subset \mathbb{C}$ and all t. For the complex-valued normal distribution and its association with the two-dimensional normal distribution we refer to Brillinger (1975). It is well known that the covariance function of X_t is given by

$$\mathcal{B}(t) = \sigma^2 e^{-\eta|t| - i\tau t}.$$

The real part of $\mathcal{B}(t)$ can now be identified with the sanogenesis curve derived above, i.e.,

$$Re\ \mathcal{B}(t) = \mathcal{S}(t).$$

To estimate the parameters η and τ, and therefore the sanogenesis curve and the associated survivor function, at each time point t a factor analysis is applied to the observed data matrix

$$\mathbf{Z}_t = [Z_{ti} : i = 1, \ldots, m]'$$

obtained from a random sample of m patients. Note that at different time points the contribution of the underlying variables to the factors may differ, since the factor analysis is applied for each time separately.

Remark 4 *It is beyond the scope of the present article to discuss carefully the problem of checking the assumptions above. However, the following arguments may be useful. Assuming a two-factor model may be restrictive, but provides a reasonable description of the disease dynamics in many real applications. The assumption of stationarity of the process is of more concern and can be checked using classic tools, e.g., based on the sample autocorrelation matrix. Analogously, the normality assumption can be checked by applying appropriate tests. The assumption that the process is Markovian is, of course, crucial, since otherwise the covariance function does not comprise the sanogenesis curve. However, if the time points where measurements are made are chosen appropriately, this assumption should be reasonable in many cases.*

3 CMNS process parameters estimation

In this section we propose point and interval estimates for the parameters η, τ, and σ^2 of the CMNS process which in turn provides estimates for the sanogensis curve and the associated survivor function. Assuming the underlying two-dimensional process driving the evolution of the disease is Gaussian allows to rely on a likelihood approach.

To construct the pointwise estimates for the parameters η, τ and σ^2 we assume that we are given m independent realizations $x_j(t) = u_j(t) + iv_j(t)$ of the CMNS process with zero expectation and the covariance function of the form

$$\mathcal{B}(t) = \sigma^2 e^{-\eta|t|-i\tau t}, \qquad \eta > 0,$$

sampled at time points $t = 1, 2, \ldots, k$. In addition, it is assumed that $u_j(t)$ and $v_j(t)$ are independent. We use the notation $x_j(l) = x_{jl}$, $u_j(l) = u_{jl}$, and $v_j(l) = v_{jl}$.

Theorem. *Let*

$$\hat{\mathcal{B}}(t) = \frac{1}{(k-t)m} \sum_{l=1}^{k-t} \sum_{j=1}^{m} x_{jl}\, x^*_{j,\,l+t} \tag{3}$$

be the sampling estimate of the CMNS process covariance function,

$$A_1 = \sum_{l=1}^{k} \sum_{j=1}^{m} x^*_{jl} x_{jl} = km\hat{\mathcal{B}}(0), \quad A_2 = \sum_{l=2}^{k-1} \sum_{j=1}^{m} x^*_{jl} x_{jl}.$$

If

$$P\Big(\Big|\frac{A_1}{k} - \frac{A_2}{k-2}\Big| > \varepsilon\Big) \to 0, \textit{with} \quad m \to \infty, \textit{for all} \quad \varepsilon > 0, \tag{4}$$

then there exist the estimates $\hat{\tau}, \hat{\eta}, \hat{\sigma}$, which are asymptotically equivalent to the maximum likelihood estimates, can be expressed in terms of the covariance function sample estimate and satisfy the equations

$$tg\hat{\tau} = \hat{\varrho}_2/\hat{\varrho}_1, \tag{5}$$

$$\hat{\eta} = -\ln\sqrt{\hat{\varrho}_1^2 + \hat{\varrho}_2^2}, \tag{6}$$

$$\hat{\sigma^2} = \hat{\mathcal{B}}(0), \tag{7}$$

where $\hat{\varrho} = \hat{\mathcal{B}}(1)/\hat{\mathcal{B}}(0), \quad \hat{\varrho} = \hat{\varrho}_1 + i\hat{\varrho}_2.$

Remark 5 *The condition (4) allows to obtain simple estimates which are asymptotically equivalent to the maximal likelihood estimates. Without that condition a solution of the normal equations of maximum likelihood method is much more complicated.*

Our approach to construct confidence regions for the parameters η,τ and for the survivor function $H(t) = G(\mathcal{S}(t))$ is based on an idea of Linnik (1950). The next lemma slightly generalizes Lemma 1 from that work to our setting.

Lemma.

Under the conditions and notation of the Theorem the distribution of the statistic $\widetilde{Q} = -Q$, where

$$Q = \frac{-A_1 - rr^*A_2 + rA_3 + r^*A_3^*}{\sigma^2(1 - rr^*)}$$

with $r = e^{-\eta - i\tau}$, does not depend on the parameters η and τ. Further, $\widetilde{Q} \sim \chi^2(mk)$ where $\chi^2(df)$ denotes the χ^2 distribution with df degress of freedom. In the following we have to discuss the cases of known and unknown scale σ separately.

3.1 The case when σ is known

Here we change the representation of Q up to the form:

$$Q = \frac{-A_1 - rr^*A_2 + rA_3 + r^*A_3^*}{\sigma^2(1 - rr^*)} =$$

$$\frac{-A_1 - e^{-2\eta}A_2 + e^{-\eta-i\tau}(ReA_3 + iImA_3) + e^{-\eta+i\tau}(ReA_3 - iImA_3)}{\sigma^2(1 - rr^*)} =$$

$$\frac{-A_1 - e^{-2\eta}A_2 + ReA_3 2e^{-\eta}cos(\tau) + iImA_3(e^{-\eta-i\tau} - e^{-\eta+i\tau})}{\sigma^2(1 - rr^*)} =$$

$$\frac{-A_1 - e^{-2\eta}A_2 + ReA_3 2e^{-\eta}cos(\tau) + 2ImA_3 e^{-\eta}sin(\tau)}{\sigma^2(1 - rr^*)} =$$

$$\frac{-A_1 - e^{-2\eta}A_2 + 2\sqrt{(ReA_3)^2 + (ImA_3)^2}e^{-\eta}cos(\tau - \phi)}{\sigma^2(1 - rr^*)},$$

where $cos(\phi) = \frac{ReA_3}{\sqrt{(ReA_3)^2+(ImA_3)^2}}$ and $sin(\phi) = -\frac{ImA_3}{\sqrt{(ReA_3)^2+(ImA_3)^2}}$.

We consider the inequalities $\alpha < \widetilde{Q} < \beta$ with respect to η and τ:

$$\alpha < -\frac{-A_1 - e^{-2\eta}A_2 + 2\sqrt{(ReA_3)^2 + (ImA_3)^2}e^{-\eta}cos(\tau - \phi)}{\sigma^2(1 - rr^*)} < \beta \qquad (8)$$

Since $\eta > 0$ and $\tau \in [0; 2\pi)$, the denominator of the fraction is greater than zero, and hence (8) is equivalent to the following system of the inequalities

$$\begin{cases} \alpha\sigma^2(1 - rr^*) < A_1 + e^{-2\eta}A_2 - 2\sqrt{(ReA_3)^2 + (ImA_3)^2}e^{-\eta}cos(\tau - \phi) \\ \beta\sigma^2(1 - rr^*) > A_1 + e^{-2\eta}A_2 - 2\sqrt{(ReA_3)^2 + (ImA_3)^2}e^{-\eta}cos(\tau - \phi). \end{cases} \qquad (9)$$

Replacing rr^* by $e^{-2\eta}$ we rearrange everything into the right-hand side and obtain

$$\begin{cases} e^{-2\eta}(A_2 + \alpha\sigma^2) - 2\sqrt{(ReA_3)^2 + (ImA_3)^2}e^{-\eta}cos(\tau - \phi) + A_1 - \alpha\sigma^2 > 0 \\ e^{-2\eta}(A_2 + \beta\sigma^2) - 2\sqrt{(ReA_3)^2 + (ImA_3)^2}e^{-\eta}cos(\tau - \phi) + A_1 - \beta\sigma^2 < 0 \end{cases} \qquad (10)$$

Next we intend to trace a two-dimensional confidence region based on this system. Notice that the parameter ϕ of the system is equal to an asymptotic likelihood estimate of the parameter τ, and the confidence interval appears symmetrical about this estimate.

3.2 The case of unknown σ

If σ is unknown and has to be estimated, the arguments of the previous subsection do not apply.

Suppose that the initial process was X_1^k. Let μ and ν be natural numbers ($k > \nu > \mu$). We divide our process into two ones, X_1^μ and X_ν^k, and define two stochastic values Q_1 and Q_2 analogous of Q, but based on different parts of the initial process (on X_1^μ and X_ν^k accordingly). When changing the sign, both statistics have a χ^2-distribution with the degrees of freedom n_1 and n_2, respectively. In addition, they are independent. Hence $T = \frac{Q_1 n_2}{Q_2 n_1}$ is distributed according to a Fisher distribution with n_1 and n_2 degrees of freedom.

Consider the solution of the following inequalities with respect to the parameters η and τ

$$\alpha < \frac{-A_{1,1} - rr^* A_{2,1} + rA_{3,1} + r^* A^*_{3,1}}{-A_{1,2} - rr^* A_{2,2} + rA_{3,2} + r^* A^*_{3,2}} < \beta. \tag{11}$$

In the case of positive denominator we obtain the following system:

$$\begin{cases} A_{1,1} - \alpha A_{1,2} + e^{-2\eta}(A_{2,1} - \alpha A_{2,2}) - \\ 2\sqrt{(ReA_{3,1} - \alpha ReA_{3,2})^2 + (ImA_{3,1} - \alpha ReA_{3,2})^2} e^{-\eta} cos(\tau - \phi) > 0 \\ A_{1,1} - \beta A_{1,2} + e^{-2\eta}(A_{2,1} - \beta A_{2,2}) - \\ 2\sqrt{(ReA_{3,1} - \beta ReA_{3,2})^2 + (ImA_{3,1} - \beta ReA_{3,2})^2} e^{-\eta} cos(\tau - \phi) < 0, \end{cases}$$

where $cos(\phi) = \frac{ReA_{3,1} - \alpha ReA_{3,2}}{\sqrt{(ReA_{3,1} - \alpha ReA_{3,2}))^2 + (ImA_{3,1} - \alpha ImA_{3,2})^2}}$
and $sin(\phi) = -\frac{ImA_{3,1} - \alpha ImA_{3,2}}{\sqrt{(ReA_{3,1} - \alpha ReA_{3,2}))^2 + (ImA_{3,1} - \alpha ImA_{3,2})^2}}$.

3.3 Computation of confidence region

No matter whether σ is known or unknown, in both cases we have to solve the system of inequalities

$$\begin{cases} c_\alpha + e^{-2\eta} a_\alpha - 2b_\alpha e^{-\eta} cos(\tau - \phi) > 0 \\ c_\beta + e^{-2\eta} a_\beta - 2b_\beta e^{-\eta} cos(\tau - \phi) < 0 \end{cases}, \tag{12}$$

where $a_\alpha, b_\alpha, c_\alpha, a_\beta, b_\beta, c_\beta$ are coefficients. The difference between the cases of known and unknown variance is that we have some additional information about the coefficients. Indeed, if σ is known we have $a_\alpha > 0$ and $a_\beta > 0$. For simplicity we shall consider the case of known σ.

Since the main coefficients are positive a confidence region for $e^{-\eta}$ is an intersection, namely $D_\eta = U_\alpha \cap U_\beta \cap (0, \infty)$, where $U_\alpha = (-\infty, t^-_\alpha) \cup (t^+_\alpha, \infty)$; $U_\beta = (t^-_\beta, t^+_\beta)$ and

$$t^\pm_\alpha = \frac{b_\alpha cos(\tau - \phi) \pm \sqrt{b^2_\alpha cos^2(\tau - \phi) - a_\alpha c_\alpha}}{a_\alpha},$$

$$t^\pm_\beta = \frac{b_\beta cos(\tau - \phi) \pm \sqrt{b^2_\beta cos^2(\tau - \phi) - a_\beta c_\beta}}{a_\beta}.$$

Therefore

$$D_\eta = (max\{0, t_\beta^-\}, min\{t_\alpha^-, t_\beta^+\}) \cup (max\{0, t_\beta^-, t_\alpha^+\}, t_\beta^+).$$

Then we trace the confidence interval for τ basing on a condition that D_η is not empty. Let V_1 denote the set of all τ such that (12) has a solution. Since (12) is a quadratic equation, we obtain

$$V_1 = \{\tau : b_\beta^2 cos^2(\tau - \phi) > c_\beta a_\beta\} = (0, arccos(\frac{\sqrt{c_\beta a_\beta}}{b_\beta}))\cup$$

$$(-arccos(\frac{\sqrt{c_\beta a_\beta}}{b_\beta}) + 2\pi, 2\pi)) \cup (arccos(-\frac{\sqrt{c_\beta a_\beta}}{b_\beta}), 2\pi - arccos(-\frac{\sqrt{c_\beta a_\beta}}{b_\beta}))$$

We, naturally, consider a case of positive free coefficient. On the second step we demand that an intersection of U_β and the positive half-line must not be empty. In other words, we obtain the following inequality

$$\frac{b_\beta cos(\tau - \phi) + \sqrt{b_\beta^2 cos^2(\tau - \phi) - a_\beta c_\beta}}{a_\beta} > 0 \Longleftrightarrow b_\beta cos(\tau - \phi) > 0. \quad (13)$$

Consequently, V_1 is contracted:

$$V_1 = (0, arccos(\frac{\sqrt{c_\beta a_\beta}}{b_\beta})) \cup (-arccos(\frac{\sqrt{c_\beta a_\beta}}{b_\beta}) + 2\pi, 2\pi)). \quad (14)$$

In the case of negative free coefficient there are no restrictions and the set V_1 is the whole interval (0,2π). Using V_1 we evidently can trace a confidence set for τ : we add ϕ to each an element of V_1 and consider the result modulo 2π. Now we shall watch the intersection of U_α, U_β and positive half-line not to be empty. It isn't empty if one of the following inequalities is true

$$t_\alpha^+ > max\{0, t_\beta^-\} \quad (15)$$

or

$$t_\alpha^+ < t_\beta^+. \quad (16)$$

4 Analysis of simulated and real data

Let us now illustrate the proposed method of analysis. We start with a simulated data set. Then we discuss two analyses of real medical data sets.

4.1 Simulated data

We simulated a CMNS process with parameters $\eta = 0.3, \quad \tau = 2.5$. Confidence intervals were calculated from the simulated trajectories using a confidence level $\beta = 0.95$. Figure 2 provides the results.

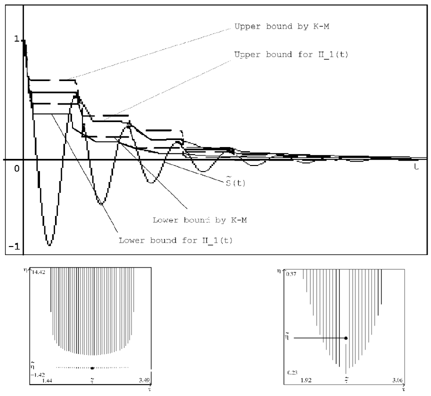

Figure 2: Confidence region for parameters and the right double inverse function of sanogenesis function; confidence level $\beta = 0.95$.

Shown are the confidence region for the parameters η and τ of the curve (left part), their estimates (black circle) and the corresponding confidence bar for $H_1(t)$ (right part; see Section 2 for the definition of $H_1(t)$). It can be seen that the confidence region for η is very asymmetrical and unbounded from above. Accordingly, it influences the confidence bar for $H_1(t)$. The latter is large and covers most of the region, lying below $H_1(t)$. Such a situation can be changed by decreasing the confidence level, increasing the sample size or by a priori considerations to narrow the bounds for the parameter η variation.

Since the first and second approach are not advisable, we propose to bound the parameter η from above using an a-priori bound. In practice this constraint corresponds to a separation of acute processes from chronic ones. Figure 2 shows a confidence region for the parameters is presented (right low), where the constraint $\eta < 0.31$ was used. We also added a Kaplan-Meier estimate calculated by simulating from the d.f. $1 - H_1(t)$ with estimated parameters.

4.2 The arterial hypertension data

This data set consists of measurements of 62 hemodynamic and biochemical characteristics og 272 patients with arterial hypertension (AH). Measurements were taken at fixed critical time points. The results of a factor analysis are given in detail in Sokolova, Bondarenko, and Evdokimova (1992). We applied the proposed procedure to estimate a sanogenesis curve from the data. Further, we mapped the sanogenesis curve to a survivor function using $\gamma = 1$ which corresponds to the most favorable disease course. AH can be classified in several variants, and the most important parameters for that task are the distance between latent time points (parameter d) and the arrangement of critical time points. The initial time point corresponds to the norm, critical ones T^1_{cr}, T^2_{cr}, T^3_{cr} — to borderline AH (BAH), AH — 1-st stage (AH1) and AH — 2-nd stage (AH2).

The patients of the study were followed up during ten years from 1981 to 1991. A factor analysis of the initial characteristics of the patients with BAH distinguished three groups according to their hemodynamic profiles depended on the patients' cardiac index (CI) and peripheral vascular resistance (PVR), cf. Sokolova, Bondarenko andEvdokimova (1992). These three groups included patients with eukinetic (a), kinetic (b) and hyperkinetic (c) hemodynamic variants. Clinical-hemodynamic prognosis of the hypokinetic patients was the worst that was reflected when interpreting the first and the third factor. We fitted sanogenesis models to cases with different kinetic types separately. The results are presented in the table 1 and figure 3

The data analysis in critical points, which were characteristic of hemodynamic group showed that the critical moments onset time is proportional to their hemodynamic level: the most early one corresponds to the hypokinetic type, and the later one to the hyperkinetic type. The same is true for the

Table 1:

Sanogenesis curve parameters estimates
by the dynamics data of arterial hypertension

Group	1-st and 2-nd factors				1-st and 3-rd factors			
	$\lvert\tau\rvert$	η	d	q	$\lvert\tau\rvert$	η	d	q
a	1.08	2.41	5.81	$1 \cdot 10^{-6}$	1.68	1.96	3.74	$2 \cdot 10^{-4}$
b	1.32	1.07	4.74	$6 \cdot 10^{-3}$	1.19	1.05	5.28	$4 \cdot 10^{-3}$
c	1.17	1.41	5.36	$5 \cdot 10^{-4}$	1.09	2.05	5.76	$7 \cdot 10^{-6}$

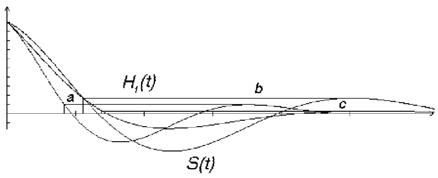

Figure 3: Sanogenesis curves for hemodynamic groups.

beginning and duration of the latent period. All this means, that *the hypokinetic group has the largest rate of AH progress.* Moreover, *in theeukinetic group the most long-lasting latent period is observed.* The survivor functions models $F_1(t)$ obtained from that of sanogenesis $S(t)$ by its double inversion are therefore interpreted in this case as curves reflecting the worsening of the miocard conteractivity.

It should be pointed out, that in the dynamics of the first two factors no analogous tendency was observed, because the variables SI and PVR contributed only tot the 1-st and the 3-rd factor but not to the 2-nd.

4.3 The surgery data

This example is of considerable interest, since the wound process is known to show a periodic pattern. A transition from one (qualitative) state to another one corresponds to one week.

The statistical analysis of data from wound processes was a part of a general study about the adequate use of medications for the correction of postaggressive adaptational reactions (PAR) in different groups of patients at the Maxillo-Facial Surgery and Stomatology Department of the Russian Military Medical Academy. 11 variables characterizing the PAR conditions were collected for 122 patients at the end of four consecutive weeks.

Additionally, the healing time, the time of out-patient, and the duration of hospitalized treatment were collected. According to the character of the disease there were 5 groups of patients. To make the idea clear we'll analyze the group of patients with maxillo-facial phlegmonas (26 pts, 9 variables which were included in test massive data).

A factor analysis revealed the following two latent factors. The 1-st factor describes the character of course of PAR, whereas the 2-nd one summarizes resuscitative and defense effects in tissues during inflammation process. Based on that analysis, the patients were classified into two groups according to the type of the course of a disease: either favorable or unfavorably. A parametric sanogenesis model was used to describe the development of the disease. In addition, Kaplan-Meier estimates of the survivor functions of the a three survival times mentioned above were calculated. The results are summarized in Figure 4. For the comparison of the model with the survivor function we used a Kolmogorov-Smirnov test, which is applicable since the parameters of the model were estimated by different observations. survivor function and hence the hypothesis of goodness of fit can be treated as a simple one. For the variable *healing time* we obtained a p value of 0.39, i.e., a good fit.

The minimal values of the healing time and out-patient treatment correspond to 20 day. By virtue of norming of a survivor function the model of development of these 2 additional variables, shown in the figure 4, can be interpreted as the conditional probability that the additional variable exceeds the argument given that its own value is not less than 20 day. As for the hospital treatment additional variable, a beginning of its breakdown should be transmitted into the second latent point (the 13-th day).

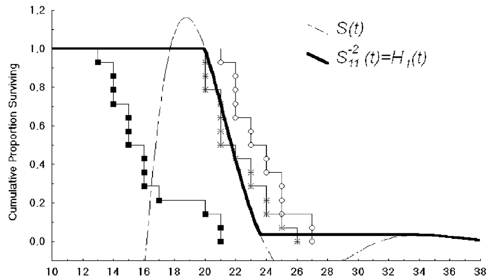

Figure 4: Estimated sanogenesis curve $\widehat{S}(t)$ (long dashed), associated survivor function $H_1(t)$ (solid line), and Kaplan-Meier estimates for the healing time (stars), out-patient treatment (circles) hospital treatment time (squares). The curves: $S(t)$ — long dash line, $H_1(t)$ — solid line.

5 Discussion and conlusions

Our approach to reveal a parametric sanogenesis curve from high-dimensional data on disease characteristics has several merits. First, the parametric form is motivated by considering interacting pathogenetic and sanogenetic factors which is a plausible model for the disease process. Second, it allows to take into account all information that is available. Third, it is possible to attach a survivor function to each sanogenesis curve. Depending on the conrete application and the meaning of the latent factors, the assumption that the survivor function derived from the sanogenesis curve can be identified with the survivor function of an appropriately defined survival time variable can be justified. That problem can be solved in a collaboration of the statistician and the practitioner, as shown by our data examples. Also note that the parametric model and the nonparametric estimate contain different information which complement each other. Further, different time variables determining survivor functions give different information in characterizing the total disease process of interest.

It should be noted that in terms of this model one can describe different forms of a disease course, since the parameterized double inversion of the sanogenesis curve allows to construct curves which describe the evolution of different variants of a disease, depending on the choice of that parameter. Assuming that the interaction takes place between an organism and a damaged organ, $\gamma = 1$ models the best course with complete control by the organism. When $0 < \gamma \leq 1$, capabilities of the organism are limited, whereas $\gamma = 0$ corresponds to the most unfavorable variant, where the state of the damaged organ completely determines the situation. The glomerulonephritis data already mentioned at the beginning of Section 2 provides an example, where a critical point separated the disease terminal stage from the rest of the process (Bart, Bondarenko, and Boiko; 1980).

A further important aspect of the method is that it models the disease development as a curve which allows to define critical time points and latent periods. The survivor function calculated from the sanogenesis curve respects that feature. Consequently, under this model one dies during crises as indicated by the sanogenesis curve.

We used factor analyses to reduce the dimensionality of the relevant data and to estimate paramters. All the information about the variables that is contained in the sanogenesis curve corresponds to the two factors. This may seem restrictive. Extracting additional information from more factors requires further research in the direction of multidimensional CMNS processes invoking other methods of multivariate statistics.

A Appendix

A.1 Definition of the partially inverse functions

Here the meaning of right inverse function to a function $y = f(x)$ is understood as an extreme right point of the left inverse image: $f_1^-(y) = \sup\{x : \ f(x) \geq y\}$. The meaning of the left inverse function is an extreme left point of the right inverse image: $f_0^-(y) = \inf\{x : \ f(x) \leq y\}$. The partiality parameter $\gamma \in (0;1)$ determines the intermediate variants $f_\gamma^-(y)$ for partially inverse functions as follows.

For each fixed y we shall denote its pre-image as $\mathcal{X}_y$. A minimal segment R_y, containing the set $\mathcal{X}_y$, we call a scope of this set, i.e.

$$R_y = (\inf \mathcal{X}_y, \sup \mathcal{X}_y) \cup \mathcal{X}_y.$$

If $0 < \mu(R_y) < \infty$, where μ is a Lebesgue measure, then to any $\gamma \in [0, 1]$ the following relation

$$\gamma = \mu\{\{x : \ x \leq z_\gamma\}|R_y\} = \mu\{\{x : \ x \leq z_\gamma\} \cap R_y\}/\mu R_y, \tag{17}$$

uniquely maps a point $z_\gamma \in \overline{R}_y$ (where $\overline{R}_y$ is the closure of R_y), and a partiality operation

$$x_\gamma = \sup\{\{x : \ x \leq z_\gamma\} \cap \mathcal{X}_y\} \tag{18}$$

defines the point $x_\gamma \in \overline{\mathcal{X}_y}$.

If $\mu(R_y) = 0$ and $\mathcal{X}_y$ is not empty, then the preimage consists of only one point x_γ which is mapped for every $\gamma \in [0, 1]$.

Since in our case at the first inversion all the preimages has a scope of a finite measure, we can say that every fixed $\gamma \in [0; 1]$ is corresponded to a function $f_\gamma^-(y) = x_\gamma$, which is called a partially inverse to $f(x)$. At the points where $f(x)$ is strictly monotone and continuous all the partially inverse functions to $f(x)$ coincide.

The partially inverse functions for a non-increasing one with parameter γ, equal to 0 or 1, are the classic extreme (left or right) inverse ones.

A.2 The proof of Theorem (from the section 3)

We construct a likelihood function

$$\mathcal{L}(u_{11}, \ldots, u_{km}, v_{11}, \ldots, v_{km} \,|\, \tau, \eta, \sigma^2) = \mathcal{L}(\,\cdot \,|\, \tau, \eta, \sigma^2),$$

which can be presented as

$$\mathcal{L}(\,\cdot \,|\, \tau, \eta, \sigma^2) = (\pi\sigma^2)^{-km}(1 - rr^*)^{-m(k-1)} \exp\{Q\},$$

where

$$Q = \frac{-A_1 - rr^* A_2 + rA_3 + r^* A_3^*}{\sigma^2(1 - rr^*)}, \tag{19}$$

$$r = e^{-\eta - i\tau}, \quad A_3 = \sum_{l=1}^{k-1} \sum_{j=1}^{m} x_{jl}^* x_{j,l+1} = m(k-1)\mathcal{B}^*(1).$$

Let us denote $\theta = \sqrt{rr^*} = e^{-2\eta}$ and

$$Z(\tau) = \mathrm{Re}A_3 \cos\tau + \mathrm{Im}A_3 \sin\tau = (rA_3 + r^* A_3^*)e^{\eta}/2.$$

Using these relations we have

$$Q = \frac{-(A_1 - 2\theta Z(\tau) + A_2\theta^2)}{\sigma^2(1-\theta^2)},$$

$$\ln \mathcal{L}(\,\cdot\,|\tau, \theta, \sigma^2) = -km \ln(\pi\sigma^2) - m(k-1)\ln(1-\theta^2) + Q.$$

Consider a system of equations

$$\frac{\partial \ln \mathcal{L}(\,\cdot\,|\tau, \theta, \sigma^2)}{\partial \tau} = 0, \tag{20}$$

$$\frac{\partial \ln \mathcal{L}(\,\cdot\,|\tau, \theta, \sigma^2)}{\partial \theta} = 0, \tag{21}$$

$$\frac{\partial \ln \mathcal{L}(\,\cdot\,|\tau, \theta, \sigma^2)}{\partial \sigma^2} = 0. \tag{22}$$

It follows from (20) that

$$tg\hat{\tau} = \frac{\mathrm{Im}A_3}{\mathrm{Re}A_3}. \tag{23}$$

So we obtain

$$Z(\hat{\tau}) = \sqrt{(\mathrm{Re}A_3)^2 + (\mathrm{Im}A_3)^2}$$

and use these equalities to solve the equations (21) and (22). ¿¿From (22) we find

$$\hat{\sigma}^2 = A_2\hat{\theta}^2 - 2\hat{\theta}Z(\hat{\tau}) + \frac{A_1}{km(1-\hat{\theta}^2)}.$$

Now we substitute this expression into (21) to get a cubic equation of $\hat{\theta}$:

$$A_2(k-1)\hat{\theta}^3 - Z(\hat{\tau})(k-2)\hat{\theta}^2 - (kA_2 + A_1) + kZ(\hat{\tau}) = 0.$$

By the condition (4) the left part of it can be asymptotically presented as a product and we receive an asymptotic equation

$$\left(\hat{\theta}^2 - \frac{k}{k-2}\right)\Big((k-1)A_2\hat{\theta} - Z(\hat{\tau})(k-2)\Big) = 0.$$

The solutions $\hat{\theta} = \pm\sqrt{k(k-2)}$ making no sense with $\eta > 0$, the estimate $\hat{\theta}$ is linear of a form

$$\hat{\theta} = Z(\hat{\tau})k/A_1(k-1), \tag{24}$$

and

$$\hat{\sigma}^2 = A_1/(km). \tag{25}$$

is an estimate for the variance σ^2. The formulas (5), (6) and (7) follow from (23), (24) and (25) because the statistics A_1 and A_3 can be expressed (3) in terms of the covariance function estimate.

A.3 The proof of Lemma (from the section 3)

Let us define the following function for Q (when $s > 0$)

$$\varphi_Q(s) = E(e^{sQ}) = \int\limits_{R^{2km}} e^{sQ} L(\cdot \mid \tau\eta\sigma^2) dw,$$

where $dw = du_1^1 \dots du_k^1 dv_1^1 \dots dv_k^1 \dots du_1^m \dots du_k^m dv_1^m \dots dv_k^m$; u_l^j and v_l^j are, accordingly, the real and imaginary components of the l-th element in the j-th realization. Using the representation for likelihood function we obtain

$$\varphi_Q(s) = (\pi\sigma^2)^{-km}(1 - rr^*)^{-m(k-1)} \int\limits_{R^{2km}} e^{(1+s)Q} dw \tag{26}$$

Let $\widetilde{u}_l^j = \sqrt{1+s}u_l^j$ and $\widetilde{v}_l^j = \sqrt{1+s}v_l^j$. We obtain:

$$\varphi_Q(s) = (\pi\sigma^2)^{-km}(1 - rr^*)^{-m(k-1)} \int\limits_{R^{2km}} e^Q \frac{1}{(1+s)^{km}} d\widetilde{w} = \frac{1}{(1+s)^{km}}, \tag{27}$$

where $d\widetilde{w}$ is a differential for the new variables, analogous to dw.

Now using the principle of analytic continuation we extend the domain of Q into the complex plain. We obtain that the characteristic function of Q is of the form

$$\psi_Q(t) = \varphi_Q(it) = \frac{1}{(1+it)^{km}}. \tag{28}$$

Therefore a stochastic value $\widetilde{Q} = -Q$ has a X_{km}^2-distribution .

References

[1] Barlow R.E.,Proschan F. (1975). *Statistical Theory of Reliability and Life testing*. New York: Holt, Rinehart and Winston.

[2] Bart A.G., Bondarenko B.B., and Boiko B.I. (1980). Mathematical analysis of the evolution of chronic glomerulonephritis. In: *Glomerulonephritis*. Eds. Riabov S.I., Leningrad: Medicine., pp. 213-225 (in Russian).

[3] Bart A.G., Klochkova N.P., Kozhanov V.M. (1993). The universal scheme of regulations in biosystems for the analysis of neuron junctions as an example. In: *Model-Oriented Data Analysis*. Eds. Muller W.G., Wynn H.P., and Zhigljavsky, A.A. Heidelberg: Physica-Verlag, pp. 167-177.

[4] Bart A.G., Alekseyeff (Klochkova) N.P. and Bochkina N.A. (2000). Partially inversion of functions for statistical modelling of regulatory systems, 355-371. In: *Advances in Stochastic Simulation Methods*, Series: Statistics for Industry and Technology. Eds. Balakrishnan N., Melas V.B., Ermakov S. Boston-Basel-Berlin: Burkhäser.

[5] Bart A.G., Alekseyeff (Klochkova) N.P. (1999). Critical periods in the survival curves, 253-257. In: *Statistical methods in clinical trials*. Eds. Zhigljavsky A.A. and Nekrutkin V.V., S-Pb.: ed.S-Pb.Univers. (in Russian).

[6] Bart A.G. (2002). *Analysis of the Medical and Biological Systems (The Inverse Functions Method)*. St.Petersburg: St.Petersburg University Press, (in Russian).

[7] Brillinger S. (1975). *International Series in Decision Processes.* New York: Holt, Rinehart and winston, Inc.

[8] Chen A., Bickel P. Efficient independent component analysis, *Technical Report 634*, Department of Statistics, UC Berkeley, 2003.

[9] Cox D.R. and Oakes D. (1984). *Analysis of Survival Data*. London - New York: Chapman and Hall.

[10] Gavrilov L.A. and Gavrilova N.S. (1991). *The Biology of the Life Longevity*. M. Nauka, 280p.(in Russian)

[11] Glantz S.A. (1994). *Primer of Biostatistics*. New York: McGraw-Hill.

[12] Hyvarinen A., Karhunen J., Oja E. (2001). *Independent Component Analysis*, New York: Wiley.

[13] Klein J.P. and Goel P.K. (1992). *Survival Analysis: State of the Art*, pp. 65-140. Dordrecht-Boston-London: Kluver Academic Publ.

[14] Linnik Yu.V. (1950). On one question of statistics of the dependent Observations. In: *Izvestia AN SSSR, Math.*,v.14,N6,pp.501-522.(in Russian)

[15] Lloyd E. and Ledermann W. (1984). *Handbook of Applicable Mathematics.*, v.YI:Statistics, part B. Chichester: A Wiley-Interscience Publication, John Wiley & Sons.

[16] Sokolova L.A., Bondarenko B.B., Evdokimova T.A. (1992). Comperative estimation of the cardiovascular reactivity in patients with the borderline hypertension and early stages of arterial hypertension, *Human Physiology*, N 1, 128–135, (in Russian).

[17] Sokolova L.A. et al. (1991). Clynical-hemodynamic characteristics of patients with the early stages of arterial hypertension, *Cardiology*,N 6, 11–13, (in Russian).