rivm
*research for*
*man and environment*

RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEU
NATIONAL INSTITUTE OF PUBLIC HEALTH AND THE ENVIRONMENT

RIVM report 260751 002

# Ageing and mortality
Results from the Zutphen-Study

RT Hoogenveen

April 2000

This investigation has been performed by order and for the account of the Board of Directors of the RIVM, within the framework of project 260751, Chronic Disease Modelling.

RIVM, P.O. Box 1, 3720 BA  Bilthoven, telephone: 31 - 30 - 274 91 11; telefax: 31 - 30 - 274 29 71

# Abstract

The link between competing death risks and the change of risk factor levels over time has been analysed using data from the Zutphen-Study and the model of Manton&Stallard (1988). The Zutphen-cohort consists of 878 men, initially with age 40-59 years, that have been followed since 1960. The model of Manton&Stallard describes the change of the risk factor levels among the individuals of a cohort taking into account mortality and the change of levels within the individuals. The model has been divided into one part on mortality and another on the risk factor level changes. The hazard function used is similar to the one used in the Cox proportional hazards model. For almost the same combinations of risk factors and causes of death significant effects have been found as in Cox analyses. However, using current instead of baseline risk factor measurement values result in smaller effects probably due to medication (for total cholesterol and systolic bloodpressure) or in larger effects probably due to a reverse causal relation (for BMI and lung cancer). The most interesting and striking results were found with respect to the risk factor changes over age. We found positive age-trends for all risk factors (although non-significant for cholesterol), whereas the results of simple regression analyses were not that clear. More specific results relate to the interactions with respect to the deterministic and random changes. The results of the analyses will be used for the further development of the chronic diseases modelling tools. That means, the refinement of the modelling of the changes of the risk factors mentioned above over age, and the relation (interaction) between these changes.

# Preface

This report describes the results of analyses on data from the Zutphen-Study within the scope of competing death risks and change of risk factor levels over time and age. In a foregoing report (Hoogenveen et al., 1993) results have been presented of Cox proportional hazards analyses. The model of Manton&Stallard (1988) enabled us to analyse the dynamic relation between change of risk factor levels and mortality.

We have made these new analyses for several reasons. Random changes are an essential aspect of the change of risk factor levels, and can be analysed using the model of Manton&Stallard. Mortality and risk factor level changes within individuals are the two processes that govern the change of the risk factor distribution of a cohort over time. These processes have also been described in demographic-epidemiological simulation models such as Prevent (Gunning-Schepers, 1988), TAM (Barendregt&Bonneux, 1998), CZM (Hoogenveen et al., 1998) and POHEM (Wolfson, 1991). The results of our analyses can be useful to further develop these types of models. For example, the modelling of the changes of the public health risk factor levels over age could be improved by including interactions between the risk factor specific changes. The Zutphen-Study is a longitudinal study over a long time period. Since 1960 approximately 900 men have been followed. Individual risk factor levels and morbidity and mortality outcome values have been registered. The study has resulted in many scientific publications so far, mainly for specific causes of death or mortality risk factors separately. In our analyses we have described a new integrative aspect.

The author thanks dr EJM Feskens, ir MGG van Genugten, dr SH Heisterkamp and EJM Veling for their contribution to the analyses, and last but not least dr ir PHM Janssen for giving ‘matrix-theoretical support’.

# Contents

# Samenvatting

De samenhang tussen concurrerende doodsoorzaken en de verandering van de niveau’s van de bijbehorende risicofactoren over de tijd is onderzocht met behulp van gegevens van de Zutphen-Studie en een model beschreven door Manton&Stallard (1988). De Zutphen-Studie bestaat uit een cohort van 878 mannen, dat is gevolgd vanaf 1960 en met beginleeftijd 40-59 jaar. Voor verschillende epidemiologische risicofactoren zijn herhaalde metingen uitgevoerd vanaf 1960, en zijn tijdstip en oorzaak van sterfte geregistreerd. Het genoemde model van Manton&Stallard beschrijft de verandering van de verdeling van de risicofactorniveau’s van een cohort over de tijd ten gevolge van enerzijds sterfte en anderzijds de veranderingen van de niveau’s binnen de individuen van het cohort. De verandering van de verdeling wordt beschreven door een zogenamde Kolmogorov-Fokker-Planck partiële differentiaal-vergelijking. De gebruikte mortaliteits hazard functie is volledig geparameteriseerd. De 'hazard ratio' term is een kwadratische regressiefunctie. De modelparameters zijn geschat met behulp van de methode van 'maximum likelihood'.

Met betrekking tot sterfte werden voor vrijwel dezelfde risicofactoren en doodsoorzaken dezelfde significante parameters gevonden als bij Cox analyses. Dat wil zeggen voor systolische bloeddruk voor totale sterfte, sterfte aan CVA en overige oorzaken, voor totaal cholesterolniveau voor totale en CHZ sterfte, voor Body Mass Index voor sterfte aan longkanker en overige oorzaken, en voor roken voor longkanker. De gevonden relatie tussen BMI en longkanker kan verklaard worden door een omgekeerd causaal verband.

Met betrekking tot de verandering van de risicofactorniveau's lieten de modelresultaten een duidelijke verandering met de leeftijd zien. Ter vergelijking uitgevoerde eenvoudige regressie-analyses vertoonden een minder eenduidig beeld. Dit resultaat bevestigde het verschil tussen individuele en populatie-veranderingen over de leeftijd. De resultaten lieten ook duidelijk de interacties zien tussen de veranderingen van de niveau’s van de risicofactoren over de tijd. We vonden een negatief verband tussen de verandering en het absolute niveau, het minst nog voor BMI. Voor elk van de risicofactoren was het verband met de niveau's voor de overige factoren gering, met name voor BMI. De gevonden toevalsveranderingen waren groot vergeleken met de 1-jaars deterministische veranderingen, met name voor bloeddruk en cholesterol. Dit resultaat kan echter vertekend zijn door 'regressie naar het gemiddelde'.

De resultaten van de analyses worden gebruikt voor de verdere ontwikkeling van de chronische ziekten modellering. Dat wil zeggen, de verbetering van de modellering van de verandering van de niveau's van de bovengenoemde risicofactoren over de leeftijd en de relatie (interactie) tussen deze veranderingen.

# Summary

In foregoing analyses the Cox proportional hazards model has been used to calculate hazard ratios for several specific causes of death and risk factors. However, there is also a reverse relation between risk factors and mortality: high risk level frequencies tend to decrease because of the related high mortality risks. The two-way relation between competing death risks and the change of risk factor levels over time has been analysed using data from the Zutphen-Study and the model of Manton&Stallard (1988). The Zutphen-cohort consists of 878 men, initially 40-59 years old, who have been followed since 1960. Several risk factors have been repeatedly measured and the times and causes of death have been registered. The model of Manton&Stallard describes the change of the distribution of the risk factor levels of a cohort over time taking into account mortality and the change of levels within individuals. The change of the population risk factor level distribution has mathematically been described by a so-called Kolmogorov-Fokker-Planck partial differential equation. The mortality hazard function used is fully parametric.The hazard ratio is a quadratic regression function of the risk factor levels. The model has been fitted by the method of maximum likelihood.

With respect to the outcome variable ’mortality’ for several risk factors and causes of death significant parameter estimates have been found. These were systolic bloodpressure for total mortality, and mortality due to CVA and other causes, total cholesterol level for total and CHD mortality, Body Mass Index for lung cancer mortality and mortality due to other causes, and smoking for lung cancer. The same combinations have been found using the Cox model, also using current risk factor values. The effect of BMI on lung cancer mortality can be explained by a reverse causal relation.

With respect to the risk factor changes we found positive changes over age for all risk factors, although non-significant for cholesterol, whereas simple regression analyses showed non-significant or even opposite age trends. For all risk factors we found that high levels tend to increase less than low levels. This interaction was smallest for BMI. For each risk factor the changes were most strongly related with the absolute values for that risk factor and less with those for the other risk factors. This result especially applied to BMI. For bloodpressure and cholesterol there was also a small dependency on BMI levels. The random changes found were large compared to the 1-year deterministic changes, especially for SBP and cholesterol. The latter result could be biased by 'regression dilution'.

The results of the analyses will be used for the further development of the 'chronic diseases modelling tools'. These are computer simulation models that are used to calculate the morbidity and mortality effects of trends in and intervention measures on public health risk factors. The analyses on ageing and mortality will be used to upgrade the modelling of the changes of the risk factor levels over age and the relation (interaction) between these risk factor specific changes.

# 1. Introduction

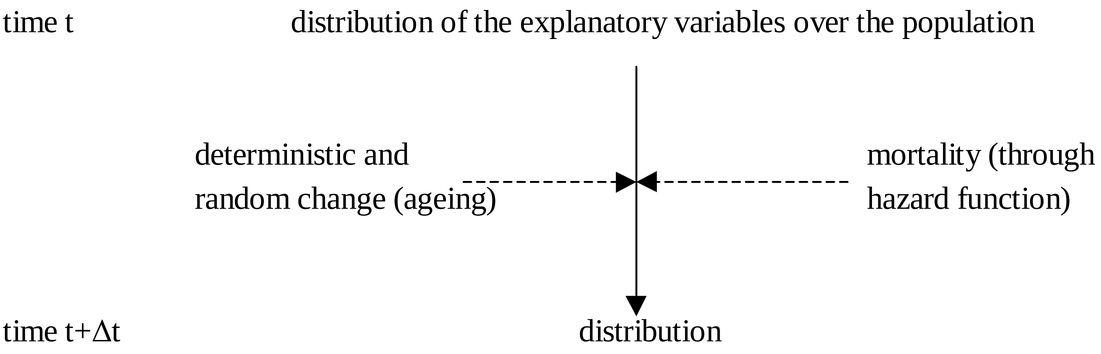
The change of the population risk factor distribution over age is the result of two processes, ageing (Mulder, 1993) and mortality. Ageing means that the risk factor levels change within individuals. This change is a stochastic process. Mortality selection means that the proportion of individuals with extreme risk factor levels and therefore with high mortality risks decrease in favour of the proportion with moderate risk factor levels and therefore small mortality risks. The model of Manton&Stallard (1988) describes the interaction of these two interrelated processes mathematically. The question we have addressed was to analyse these two processes ad their interaction using data from the Zutphen-Study.

We have presented the conceptual model in §2.1. The starting mathematical model is a so-called partial differential equation (§2.3). Making some specific assumptions the model has been simplified to model equations on the two parameters of a multivariate normal distribution (§2.4). In chapter 3 the data have been described that have been used to fit the model. These data were from the Zutphen-Study (Feskens, 1991). The model results have been presented in chapter 4. To prevent non-mathematicians from getting stuck in chapter 2, the results have been preceded by a summary of the model. We have presented results for the two parts of the model, i.e. for the part describing the risk factor level changes (§4.2) and the part describing the mortality (§4.5). In chapter 5 the results have been summarised and conclusions were drawn. In Appendix A mathematical proofs of some model development steps have been presented.

# 2. Model

## 2.1 The conceptual model

The model of Manton&Stallard (1988) describes the processes of risk factor changes and mortality and their interaction on both the individual and population level. The main characteristics of the model are the following. A multivariate function describes the joint distribution of the risk factors in a cohort. This joint distribution changes due to two processes. One process is change within individuals. Deterministic changes (drift) and random changes (diffusion) are distinguished (§2.3). The other process is mortality selection, meaning that the extreme risk factors, i.e. those with high mortality risks, disappear in expectation. The mortality hazard function is functionally dependent on these risk factors (§2.4). Different causes of death can be distinguished (see §2.6). The model structure is presented in the next scheme:



time t

distribution of the explanatory variables over the population

deterministic and
random change (ageing)

mortality (through
hazard function)

time t+Δt

distribution

## 2.2 The mathematical symbols used

**indices, numbers, etc.**

$I$ set of individuals, with index i

$I^*$ set of observed subject-intervals; each individual generates one subject-interval for each time interval being observed, with index i

$K$ set of death risks to be distinguished, with index k

$M$ set of individuals having died; also the number of individuals

$M_k$ set of individuals with cause of death k; also the number of individuals; $M = \Sigma_k M_k$

$J$ set of variables (explanatory variables for outcome mortality, epidemiological risk-factors) having been distinguished, with index j

$N$ number of time measurements on the variables, with index n

**study variables**

$t_i$ observed time of death of individual i

$C_i$ observed cause of death of individual i

$d_n$ n-th measurement time point

$x_i(t_0,t_N) = \{ x_{i0},..,x_{iN} \}$ observed variable values $x_{in}$ for individual i on time measurements $d_n$

**model variables**

$\tau$ stochastic time of death with possible value t

C cause of death

$\zeta$ stochastic explanatory variables with possible value z

$f_t(z)$ probability density function on time point t conditional on survival

$m_t,V_t$ mean and variance of the (assumed) multivariate normal probability distribution function of the explanatory variables on time point t

$u_t(z)$ deterministic change (drift) of the variable values

$\mu(t,z)$, $\mu_k(t,z)$ hazard function on time point t conditional on z for death to all causes and to death risk k respectively

$\mu(t)$ population hazard function on time t

**model parameters**

A,D parameters of the deterministic and random change respectively of the variables over time; D is assumed upper-triangular; A and D may be time-dependent

$\Sigma$ variance-covariance matrix of the random change of the variable values; $\Sigma \equiv D^TD$

$Q,Q_k$ matrices to describe the relationship between the variable $z_t$ and the hazard function $\mu(t,z_t)$ and cause-specific hazard function $\mu_k(t,z_t)$ respectively; Q and $Q_k$ may be time-dependent

$U,U_k$ unique reparameterisation of matrix Q and $Q_k$ respectively; U and $U_k$ are upper-triangular; $U^TU = Q$, $U_k^TU_k = Q_k$

The main links between the study variables and their model counterparts are:

| | model | study |
|---|---|---|
| time and cause of death | $\tau$, C | $t_i$, $C_i$ |
| explanatory variables (risk factors) | $\zeta_t$ | $x_{it}$ |

## 2.3 The Kolmogorov-Fokker-Planck partial differential equation

The variables of the model are the time and cause of death and the explanatory variables (epidemiological risk factors). The outcome variable used in this paragraph is total mortality; specification of cause of death is introduced in §2.6. The time of death is described by the hazard function. The distribution of the explanatory variables is described by a time-dependent probability density function. The hazard function and the density function are mathematically defined as:

$$\mu(t,z) = \lim_{\Delta t \downarrow 0} \Pr(\tau \leq t+\Delta t | \tau > t,z) / \Delta t$$

$$f_t(z) = \delta^J/\delta z_1..\delta z_J \Pr(\zeta_{1t} \leq z_1,..,\zeta_{Jt} \leq z_J)$$

respectively, with: $\tau$: time of death with value t; $\zeta$: vector of explanatory variables with value z; $f_t(z)$: the probability density function of the explanatory variables; $\mu(t,z)$: the hazard function. Both the probability density function and the mortality hazard function are conditional on survival until time t and the explanatory variable z. They can be linked through a partial differential equation, that can be derived using two assumptions:

1 The deterministic change of the variables is linear.
2 The random change of the variables is a so-called multivariate Wiener process

The resulting so-called Kolmogorov-Fokker-Planck partial differential equation describes the change of the distribution of the explanatory variables within the cohort over time:

$$\delta/\delta t\ f_t(z) = -\ \Sigma_j \{ \delta/\delta z_j [ u_j(z)*f_t(z) ] \} + ½\ \Sigma_{i,j} \{ \delta^2/\delta z_i \delta z_j [ \sigma_{ij} f_t(z) ] \} - \{ \mu(t,z) - \mu(t) \} f_t(z)$$

with: $u_j(z)$: the j-th component of the drift (deterministic change) vector of the variable z; $\sigma_{ij}$: the ij-th element of the covariance matrix that governs the random change of z. Note that $u_j(z)$ and $\sigma_{ij}$ may be time-dependent. For reasons of notational convenience this time-dependency has been omitted here. The right-hand side of the differential equation consists of three terms. The sum of the first two terms is called the forward diffusion operator of the process of changing variables. The first term describes the change due to drift (deterministic change). The second term describes the change due to diffusion (random change). The third term describes the change of the density function due to mortality.

The assumptions underlying the equation can be explained as follows. We assume that the change of the variable $z_t$ within any individual can be described by a linear stochastic differential equation:

$$dz_t = \{ A_0 + A_1 z_t \} dt + D^T dw_t$$

with: $A_0$: the autonomous, constant deterministic change; $A_1$: the regression coefficients that describe the linear deterministic changes; $w_t$: a so-called multivariate Wiener process with independent increments in non-overlapping time-intervals; D: matrix of scale vectors, that are independent of $z_t$. The characteristics of the Wiener process are: E(dwt)= 0, $\mathrm{var}(dw_t) = I\, dt$, with: I: the unit (diagonal) matrix. The deterministic component of the change can be described alternatively using some new notation: $A_0 + A_1 z_t \equiv A\, z_t^*$, with: $A = [\, A_0\ A_1 \,]$, $z_t^{*T} = [\, 1\ z_t^T \,]$. These variables are called extended (augmented) with respect to their constituents. The matrix D is the unique upper-triangular matrix square root of the variance-covariance matrix Σ: $\Sigma \equiv (\sigma_{ij}) = D^T D$

## 2.4 Analytical solution of the partial differential equation

The partial differential equation cannot be solved analytically, because $f_t(z)$ is not a specified parametric family. Woodbury&Manton (1977) have shown that a closed parametric family can be generated by combining three assumptions:

1 A linear dynamics model of the variable vector
2 A quadratic form of the hazard regression model
3 The explanatory variables are initially multivariate normally distributed

The first assumption already underlies the partial differential equation described above. The second assumption can be stated mathematically as follows:

$$\mu(t,z_t) = z_t^{*T}\, Q\, z_t^*$$

with: $z_t^*$: the augmented variable (see above); Q: a square symmetric non-negative definite matrix. The matrix Q can be partitioned at the first row and column to isolate the constant, linear, and quadratic coefficients. The multiplier ½ is similar to the one used in the mathematical definition of the normal distribution. Thus:

$$Q = \begin{vmatrix} b_0\ \tfrac{1}{2}b_1^T \\ \\ \tfrac{1}{2}b_1\ \tfrac{1}{2}B \end{vmatrix}$$

$$\mu(t,z_t) = z_t^{*T}\, Q\, z_t^* = b_0 + b_1^T\, z_t + \tfrac{1}{2}\, z_t^T\, B\, z_t$$

The quadratic form of the hazard function (U shape) makes sense epidemiologically. E.g. for BMI quadratic models have been used before to describe the relation with total mortality (Menotti, 1996). Under these three assumptions z is for all time points also multivariate normally distributed, with the distribution denoted by $N(m_t,V_t)$. $m_t$ is the mean, $V_t$ is the covariance matrix. $m_t$ and $V_t$ satisfy the following set of ordinary differential equations:

$$d/dt\ m_t = \{ A_0 + A_1 m_t \} - V_t \{ b_1 + B\ m_t \}$$

$$d/dt\ V_t = \Sigma + V_t A_1^T + A_1 V_t - V_t B V_t$$

The change of the mean value can be partitioned into two terms. The first term is identical to the deterministic change according to the linear dynamics model. The second part describes the effect of mortality selection. The change of the covariance matrix is described by four terms. The first term describes the variance due to the random change. The next two terms describe how the deterministic change influences the covariance matrix. The last term describes the effect of mortality selection. The mean and variance of the population hazard function are:

$$\mu(t) = E( \mu(t,z_t)|\tau>t ) = \mu(t,m_t) + ½\ \mathrm{trace}( V_t B )$$

$$\mathrm{var}( \mu(t,z_t)|\tau>t ) = m_t^T B V_t B m_t + ½\ \mathrm{trace}( B V_t B V_t ) + 2 m_t^T B V_t b_1 + b_1^T V b_1$$

$V_t B$ is a positive definite matrix, which has a positive trace value. The first equation shows that the mean mortality hazard function of a population is always greater than the hazard function evaluated at the mean variable values. This result is well-known: neglecting population heterogeneity always results in overestimating individual hazard rates.

The hazard function used is similar to the Cox proportional hazards model. It is the product of a baseline hazard function and a hazard ratio. However, the functional forms of the hazard ratios are different, quadratic instead of loglinear. We have analysed two parameterisations of the Cox model, one assuming proportional cause-specific baseline hazard functions (using extra proportionality coefficients), and one without this assumption. The cause-specific mortality hazard rates used are:

| | |
|---|---|
| proportional hazards | $\mu_k(t,z) = \exp(\alpha_k+ ß_0 t+ß_k'z)$ |
| non-proportional hazards | $\mu_k(t,z) = \exp(\alpha_k+ ß_{0k} t+ß_k'z)$ |

respectively, with: $\mu_k(t,z)$: cause-specific mortality hazard function, t: time, z: variable vector, $\exp(ß_k'z)$: hazard ratio, $ß_k$: regression coefficients with respect to all explanatory variables (except age), $ß_0$, $ß_{0k}$: regression coefficient with respect to age, $\alpha_k$: cause-specific proportionality coefficient.

Because the functional forms of the hazard ratios are different, the regression parameters of both models cannot be compared easily. To simplify making comparisons we have fitted a linear instead of quadratic hazard model, that can be interpreted as the first order approximation to the loglinear model. When using a linear hazard ratio model, the hazard function may become negative and some specific second order characteristics of the model may get lost. In case of a linear regression model, the hazard functions for two individuals with equal values for all explanatory variables except for one unit difference for variable k are:

| | |
|---|---|
| explanatory variable values: | $x_{2t} = x_{1t} + e_k$ |
| Cox model: | $h_i(t;x_t) = h_0(t) \exp( ß' x_{it} )$ |
| Manton&Stallard model: | $h_i(t;x_t) = h_0(t)\, ß' x_{it}$ |

with: i=1,2: index for the individuals. Then the hazard ratios $h_2(t)/h_1(t)$ become:

| | |
|---|---|
| Cox model: | $\exp( ß' x_{2t} ) / \exp( ß' x_{1t} ) = \exp( ß_k ) \approx 1 + ß_k$ |
| Manton&Stallard model: | $ß' x_{2t} / ß' x_{1t} = 1 + ß_k / ß' x_{1t}$ |

These ratios show the main difference between both hazard functions. The Cox hazard function is fully multiplicative. The linearised version of the Manton&Stallard hazard function is partly multiplicative (hazard ratio and baseline risk) and partly additive (effects of the explanatory variables described within the hazard ratio).

For computational reasons the continuous-time model has been approximated by a discrete-time model. That means, the differential equations that describe the instantaneous change of the mean and variance-covariance matrix have been transformed to discrete-time equations. These equations describe the mean and covariance values at the end of a small discrete time step given the values at the start. First the effects of mortality are calculated over one time step, assuming no change of the variables within individuals. Next the effects of ageing are calculated, assuming no mortality. The mathematical equations of the discrete-time model are given below, assuming a unit time step.

**Step 1: effects of mortality, assuming constant variables**

$$V_{t+1} = D_t V_t$$

$$m_{t+1} = D_t ( m_t - V_t b_1 ) = m_t - D_t V_t (b_1 + B m_t)$$

$$\mu_t = |D_t|^{½} \exp( - b_0 - b_1^T D_t m_t + ½\, b_1^T D_t V_t b_1 - ½\, m_t B D_t m_t )$$

with: $m_t$, $m_{t+1}$, $V_t$, $V_{t+1}$: the mean and variance coefficient of the multivariate normal distribution of the variable z on time t and t+1 respectively; $D_t = ( I + V_t B )^{-1}$.

**Step 2: effects of ageing, assuming no mortality**

$$m_{t+1} = A_0 + ( I + A_1 )\, m_t$$

$$V_{t+1} = \Sigma + ( I+A_1 )\, V_t\, ( I+A_1^T)$$

In these equations it is implicitly assumed that the discrete steps start from the same time point. In the model implementation (see also Manton&Stallard, 1988) the two steps have been ordered. That means, the second step starts where the first step has ended.

## 2.5 The likelihood function

The model parameters have been estimated by maximum likelihood. The likelihood function used is based on the discrete-time approximate model version:

$$L( \{x_i(t_0..t_N)\},\{t_i\}_{i\varepsilon I} ; A,\Sigma,Q ) \propto$$

$$\Pi_{i\varepsilon I} \quad f_0(x_{i0})\ \Pi_{0\leq n<[ti]}\ f_n(x_{i,n+1}) \mid x_{in} ) \ * \qquad (L_D)$$

$$\Pi_{0\leq n<[ti]} \exp\{ -\mu( d_n,x_{in} ) \} \exp\{ -\Theta_i\, \mu( [t_i],x_i([t_i]) ) \} \ * \qquad (L_M)$$

$$\mu( [t_i],x_i([t_i]) )$$

with (see also §2.2):

$i\in I$ index over individuals  
$n$ index over time measurement points  
$d_n$ n-th measurement time point  
$f_t(x_{t+1}|x_t)$ the multivariate probability density function of the explanatory variables x on time t+1 conditional on the values on time t; f is the multivariate normal distribution  
$\mu(t,x_t)$ the hazard function during time period t conditional on x

**study variables (data):**

$t_i$ the time of death of individual i  
$[t_i]$ the 'floor' value of $t_i$; i.e. the greatest integer value that is not greater than $t_i$  
$\Theta_i$ the fractional part of the time of death; $\Theta_i = t_i-[t_i]$  
$x_i(t)$ observed values of the explanatory variables on time point t for individual i

**model parameters:** A, Σ and Q

The likelihood function has been split up in two parts, with non-overlapping model parameters. That means, to fit the total model the submodels on mortality and risk factor changes have been fitted independently. The 1st part (1st line) of the likelihood function describes the joint probability density function of the variables over time, and is denoted by $L_D$. The joint probability function is split into series of conditional density functions. That means, the initial variable values at the start of the simulation period, and for all further time intervals the values at the end conditional on the values at the start:

$$L_D \propto \Pi_{i\varepsilon I} \{ f_0(x_{i0}) \; \Pi_{0<n\leq[\tau i]} f_n( x_{in} \mid x_{i,n-1} ) \}$$

The 2nd part (2nd and 3rd line) of the likelihood function describes the probability function of the times of death, and is denoted by $L_M$. The probability of dying is equal to the survival probability times the mortality hazard rate. The probability of survival is split up into conditional survival probabilities over successive time intervals:

$$L_M \approx \Pi_{i\varepsilon I} \{ \Pi_{0\leq n<[\tau i]} \exp\{ -\mu( t_n,x_{in} ) \} \;\; \exp\{ -\Theta_i \; \mu( \tau_i,x_i([\tau_i] ) ) \} \;\; \mu( [\tau_i],x_i([\tau i]) )$$

Both parts of the likelihood function are built up from congruent terms for each observation time interval for each individual. These intervals are called subject-intervals. Using this concept the likelihood function can be reformulated as, omitting the likelihood of the initial values:

$$L \propto \Pi_{i\varepsilon I^*} f_n( x_{ie} \mid x_{ib} ) \; \Pi_{i\varepsilon I^*} \exp( -w_i \;\; \mu_i ) \;\; \Pi_{i\varepsilon M^*} \mu_i$$

with:

$I^*$ set of subject-intervals being observed, again with index i
$M^*$ set of subject-intervals in which the relating individual dies
$x_{ib}$,$x_{ie}$ risk factor values at start and end respectively of subject-interval
$w_i$ length of subject-interval i
  $w_i$ = 1 if individual is being observed during full time period
  $\Theta_i$ if individual dies during the time period
  0 otherwise
$\mu_i$ hazard rate for subject-interval i; $\mu_i = x_{ib}{}^{*T} \; Q \; x_{ib}{}^*$

For each subject-interval conditional on the risk factor values at the start, the difference between the values at the end and start is multivariately normally distributed:
$x_{ie}-x_{ib}|x_{ib} \propto N(A_0+A_1x_{ib},\Sigma)$, with: $A_0$: the vector of constant risk factor changes; $A_1$: the matrix of changes proportional to the absolute values; Σ: the varaince-covariance matrix. The loglikelihood function for any subject-interval is (omitting the constant):

$$\ln f(x_{ie}|x_{ib};A,\Sigma) = - \tfrac{1}{2}\ln| \Sigma | -\tfrac{1}{2}(x_{ie}-\mu)' \; \Sigma_t^{-1}(x_{ie}-\mu)$$

with: μ, Σ : the mean value and covariance matrix of the multivariate normal distribution. The parameters are defined as: $\mu = A_0 + A_1 \; x_{ib}$, $\Sigma = D^T \; D$. The determinant |Σ| is equal to the product of the eigen-values of Σ.

## 2.6 Distinguishing different causes of death

The ‘risk factor change submodel’ and the relating part of the likelihood function do not change when distinguishing different causes of death. However, the ‘mortality submodel’ does change. We assume that all cause-specific mortality hazard functions are functionally dependent on the explanatory variables through the same quadratic functions:

$$\mu_k(t,z_t) = z_t^{*T} Q_k z_t^*$$

with: $z_t^*$: the augmented explanatory variable (see before); $Q_k$: symmetric non-negative definite matrix. Then the total mortality hazard function is functionally dependent on the explanatory variables following the same quadratic model, the matrix Q being the sum of the cause-specific matrices $Q_k$:

$$\mu(t,z_t) = \Sigma_k \mu_k(t,z_t) \qquad => \qquad Q = \Sigma_k Q_k$$

The part of the likelihood function related to mortality becomes, using again the concept of subject-intervals:

$$L_M \propto \Pi_{k \varepsilon K} \{ \Pi_{i \varepsilon I*} \exp\{ -w_i \mu_{ik} \} * \Pi_{i \varepsilon Mk*} \mu_{ik} \}$$

with: K set of causes of death being distinguished with index k

$\mu_{ik}$ cause-specific mortality hazard function value during i-th subject-interval

$M_k^*$ set of subject-intervals in which individuals die due to cause of death k

So the mortality part of the likelihood function can be separated into equivalently structured cause-specific terms. This property is called multiplicative separability. The maximum likelihood estimates of the regression coefficients can be found by maximising the cause-specific terms separately.

## 2.7 Autonomous changes over time

The quadratic hazard regression model can be defined conditional on an autonomous change over time. Manton&Stallard (1988) describe a two-stage estimation strategy to estimate both the autonomous change and regression parameters. The first stage involves the estimation of the parameters of the regression model conditional on the autonomous time trend. The second stage involves the estimation of this trend parameter. The authors use a Gompertz-type function that describes an exponential increase over trend:

$$h(t) = C_1 \exp( C_2 t )$$

Instead of representing the parameters $C_1$ and $C_2$ explicitly in the likelihood function they have employed a data transformation. For the Gompertz function this data transformation has the following form:

$$x_t \leftarrow x_t \exp(\text{ß}(age(t_0)+t+0.5-t_0-age_0) / 2)$$

with: $t_0$: the year for which the estimated quadratic hazard coefficients apply directly (1960), $age(t_0)$: the age on $t_0$, $age_0$: the age for which the estimated quadratic hazard coefficients apply directly. After substituting the data transformation in the hazard function the quadratic hazard regression model becomes:

$$h(t;x) = \{ x_t^T \exp(\tfrac{1}{2}\text{ß}(t+0.5-t_0)) \} \, Q \, \{ x_t \exp(\tfrac{1}{2}\text{ß}(t+0.5-t_0)) \} =$$

$$x_t^T \, Q \, x_t \exp(\text{ß}(t+0.5-t_0))$$

This is the desired exponential form of the hazard function. In case of a linear hazard function (see §4.5) the term ½ is omitted in the data transformation formula.

## 2.8 Including missing values and multiple unit time steps

Two types of missing values of explanatory variables are found in the Zutphen-Study data. One type is due to right-censoring because of mortality. The other type is truly missing. For bloodpressure and smoking no measurements have been made at specific time points. We have fitted the ‘risk factor change submodel’ on the time period during which measurements were made for all time points for all risk factors except for smoking, i.e. 1960-1970. In this way the information from all risk factor measurements in 1977 and 1985 is not used. We have analysed how the model results would change when including data for 1977. For bloodpressure all data in 1997 are missing, and so we have used imputed values. Including data for 1985 is not useful, because the time- and age-range would be too different from the ranges based on the time period 1960-1977.

The first step to impute missing values was to fit a linear regression model for each risk factor with all other factors (including age) as covariables. These models have been used to 'predict' the missing values. We have sampled from distributions instead of imputing the expected values to maintain the variability structure of the imputed data. The distributions used are based on the following theorem on multivariate normal distributions (Muirhead, 1982): let x be multivariately normally distributed $N_m(\mu,\Sigma)$ with x, μ, and Σ be partitioned as:

$$x = \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} , \qquad \mu = \begin{vmatrix} \mu_1 \\ \mu_2 \end{vmatrix} , \qquad \Sigma = \begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}$$

then the conditional distribution of $x_1$ given $x_2$ is multivariately normal with expectation $E(x_1|x_2) = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2-\mu_2)$ and covariance $var(x_1|x_2) = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$.

The imputation of missing systolic bloodpressure values for the year 1977 results in a new data set with multiple unit time steps, e.g. length 7 between 1970 and 1977. In first order approximation the deterministic and random changes over n time steps are equal to n times the change over one unit time step. For longer time intervals this first order approximation is too crude. To describe the changes over n steps we have recursively applied the formula for one unit time step. The resulting change of the risk factors over n time steps is:

$$z_n - z_0 = \Sigma_{i=0}^{n-1} A_1'^{i} A_0 + (A_1'^{n}-I)\, z_0 + \Sigma_{i=0}^{n-1} A'^{i} D^T w_{n-1-i}$$

with: $z_0,z_n$ initial risk factors and those after n time steps respectively;
$z_{n+1} \equiv z_n + A_0 + A_1 z_n$
$A_0,A_1$ vector of constant changes and matrix of proportional changes respectively;
$A_1' \equiv I+A_1$
I unit diagonal matrix

The expected values and variances can directly be calculated from this formula:

$$E(z_n-z_0|z_0) = \Sigma_{i=0}^{n-1} A_1'^{i} A_0 + (A_1'^{n}-I)\, z_0$$

$$Var(z_n-z_0|z_0) = \Sigma_{i=0}^{n-1} A_1'^{i}\, \Sigma\, (A_1')^{T\ i}$$

## 2.9 Comparison to demographic-epidemiological simulation models

The model of Manton&Stallard gives a simultaneous description of the processes of risk factor level changes and mortality for a cohort. It has been compared to other demographic-epidemiological computer simulation models that have been developed inside and outside the Netherlands. Examples of these simulation models are Prevent (Gunning-Schepers, 1988), TAM (Barendregt&Bonneux, 1994), CZM (Hoogenveen et al., 1998), POHEM (Wolfson, 1991). These public health models can be characterised in the following way:

a They describe the population distributed over several risk factor classes, specified by gender and age. Most often also disease morbidity is included in the model by describing disease prevalence incidence and prevalence numbers.

b The models are system-dynamic. I.e. persons can move from one state to another. These transitions are used to describe the processes of ageing, change of risk factor class, disease incidence, disease progress and mortality.

c The models are non-parametric. I.e. all distributions (over diseases states for each disease category, over classes for each risk factor) are non-parametric.

d The mathematical models used are stochastic difference equations. As long as all transition rates are independent on the state population numbers, the model results can be interpreted as mean population numbers.

e The models are Markovian: conditional on the actual population numbers in the model states, the future numbers are independent on the past numbers.

For each aspect we describe the related characteristic of the model of Manton&Stallard:

a It describes continuous risk factor levels of a cohort, specified by gender and age. The model does not describe disease morbidity.

b The model is also system-dynamic.

c The model is parametric: the risk factor levels are assumed to be multivariate normally distributed.

d The model is also stochastic. All distributional characteristics are known.

e The model is also Markovian.

We conclude that the model of Manton&Stallard is in many aspects similar to the simulation models being mentioned, although there are also some major differences. In its actual form it can only be applied complementary to these simulation models; direct integration is too complicated.

# 3. Data

## 3.1 Introduction

The data that have been used to fit the model are from the Zutphen-Study (Feskens, 1991; Kromhout et al., 1982; Voedingsraad, 1984). The Zutphen-Study is a longitudinal cohort study. It has been started in 1960 on 878 men from the Dutch town Zutphen who were born between 1900 and 1920. The Zutphen-Study is the Dutch contribution to the Seven Countries Study, that has been initiated by Keys (1980). The cohort consists of a random sample from 2 out of 3 that has been drawn from the Zutphen population registry after stratification into five-year age classes. In the year 1960 the participants were aged 40 to 59 years. A great number of epidemiological risk factors have been measured on the individuals in this year. Repeated measurements on several risk factors have been made during the following years. The mortality follow-up has been closed in 1990. The time and cause of death of all individuals have been registered. Also incidence data for some diseases have been measured.

The risk factors that have been used in our analyses are:

- age at the start of the Zutphen-Study (i.e. 1960),
- systolic bloodpressure (abbrevation: SBP, unit: mmHg),
- serum cholesterol level (abbr.: chol, unit: mg/dl),
- number of cigarette-years, i.e. the product of the number of cigarettes smoked per year times the smoking period in years (abbr.: sigyr),
- Body Mass Index that measures the relative weight of persons (abbr.: BMI, unit: $kg/m^2$).

## 3.2 Data summaries

Several characteristics have been presented of the risk factor distributions, i.e. the minimum, maximum, median and mean values (see Table 1-4), the correlations with time and age (see Table 5), and finally scatter plots and Box plots (Figure 2-22), that show the relation of the risk factor values with time and age.

*Table 1 Systolic bloodpressure (mmHg)*

| year | min | max | med | mean | #NA's |
|---|---|---|---|---|---|
| 1960 | 100 | 250 | 140 | 143 | 6 |
| 1961 | 100 | 240 | 138 | 139 | 75 |
| 1962 | 105 | 235 | 140 | 142 | 102 |
| 1963 | 98 | 224 | 134 | 140 | 133 |
| 1964 | 90 | 216 | 134 | 136 | 153 |
| 1965 | 98 | 230 | 140 | 142 | 161 |
| 1966 | 100 | 245 | 140 | 144 | 199 |
| 1967 | 110 | 260 | 150 | 151 | 215 |
| 1968 | 100 | 240 | 148 | 150 | 220 |
| 1969 | 100 | 240 | 150 | 149 | 226 |
| 1970 | 100 | 230 | 144 | 147 | 255 |
| 1977 | - |  |  |  |  |
| 1985 | 105 | 215 | 149 | 150 | 518 |

*Table 2 Serum cholesterol level (mmol/l)*

| year | min | max | med | mean | #NA's |
|---|---|---|---|---|---|
| 1960 | 2.40 | 12.62 | 5.97 | 6.10 | 50 |
| 1961 | 2.17 | 10.32 | 5.97 | 6.08 | 84 |
| 1962 | 2.30 | 10.86 | 6.34 | 6.44 | 101 |
| 1963 | 1.84 | 15.28 | 5.77 | 5.90 | 123 |
| 1964 | 2.07 | 9.34 | 5.77 | 5.90 | 148 |
| 1965 | 2.15 | 9.98 | 5.95 | 6.03 | 160 |
| 1966 | 2.20 | 10.47 | 6.21 | 6.34 | 193 |
| 1967 | 2.59 | 9.39 | 6.10 | 6.15 | 199 |
| 1968 | 2.48 | 11.92 | 6.28 | 6.34 | 219 |
| 1969 | 2.59 | 10.24 | 6.21 | 6.23 | 222 |
| 1970 | 2.30 | 10.42 | 6.15 | 6.18 | 254 |
| 1977 | 3.26 | 10.45 | 5.82 | 5.90 | 308 |
| 1985 | 2.84 | 13.21 | 6.13 | 6.15 | 517 |

*Table 3 Body Mass Index (kg/m$^2$)*

| year | min | max | med | mean | #NA's |
|---|---|---|---|---|---|
| 1960 | 16.6 | 36.6 | 24.1 | 24.1 | 3 |
| 1961 | 16.8 | 36.6 | 24.8 | 24.8 | 160 |
| 1962 | 17.8 | 36.6 | 24.7 | 24.8 | 106 |
| 1963 | 17.8 | 36.6 | 24.7 | 24.7 | 129 |
| 1964 | 17.8 | 36.6 | 24.9 | 24.8 | 151 |
| 1965 | 17.9 | 36.6 | 24.9 | 25.0 | 161 |
| 1966 | 17.8 | 36.6 | 25.1 | 25.2 | 192 |
| 1967 | 18.2 | 36.6 | 25.2 | 25.3 | 201 |
| 1968 | 17.9 | 36.6 | 25.3 | 25.3 | 219 |
| 1969 | 17.1 | 36.6 | 25.3 | 25.3 | 226 |
| 1970 | 17.1 | 36.6 | 25.2 | 25.3 | 257 |
| 1977 | 17.0 | 36.6 | 24.9 | 25.1 | 305 |
| 1985 | 15.4 | 37.4 | 25.2 | 25.4 | 518 |

*Table 4 Number of cigarette years*

| year | min | max | med | mean | #NA's |
|---|---|---|---|---|---|
| 1960 | 0 | 1575 | 353 | 373 | 14 |
| 1965 | 0 | 1650 | 405 | 418 | 109 |
| 1970 | 0 | 1775 | 435 | 455 | 187 |
| 1977 | 0 | 1705 | 480 | 500 | 308 |
| 1985 | 0 | 3060 | 456 | 518 | 527 |

Only small changes of the characteristics over time have been found. The medians and means of the risk factor distributions do not differ much, except for the number of cigarette years. This points at approximately symmetrical distributions, and we have used in the analyses the risk factor data (not including cigarette years) without any transformation. The number of missing values increases over time. The differences between the numbers of missing values for the risk factors become smaller. The main reason for missing values is right censoring due to mortality. The number of cigarette years and the serum cholesterol level show the greatest variation within the population.

In Table 5 the correlations between the risk factors and time and age have been presented. For every combination two numbers have been presented: the number above is the linear correlation based on only the measurement values in the starting year (1960), the number below is the linear correlation based on all measurement values. Missing values have been treated as missing at random.

*Table 5 The linear correlations between the risk factors and with time and age*

| | **chol** | **sigyr** | **BMI** | **time** | **age** |
|---|---|---|---|---|---|
| **SBP** | .114 | .006 | .324 | | .123 |
| | .108 | .037 | .309 | .160 | .216 |
| **chol** | | .069 | .210 | | -.027 |
| | | .044 | .171 | -.006 | -.062 |
| **sigyr** | | | -.007 | | .090 |
| | | | .015 | .148 | .136 |
| **BMI** | | | | | -.002 |
| | | | | .089 | 0.027 |

The risk factors that have been distinguished do not correlate much mutually, except for BMI (but not with cigarette years). SBP and cigarette years correlated positively with time and age. Cholesterol values seemed to decrease over time and age. The age-trend of BMI was not clear from the data. The correlations of the risk factor values with age only in 1960 were smaller (in absolute value) than those with age for all measurement time points. One is tempted to draw simple conclusions from these correlations. However, this is not allowed. The main reason is that missing values are not random: high risk factor levels have larger probabilities of being missing due to mortality.

We have also presented some scatter plots and Box plots that show the relation between the risk factors and time or age. In these figures the systolic bloodpressure, serum cholesterol, and Body Mass Index level, and the number of cigarette-years measurement values have been plotted against time and age, and also the differences between two successive measurement values for all individuals. The differences have only been plotted for the years between 1960 and 1970 because only during this period the measurement time points are equally spaced.

The scatter plots cannot be used to make definite conclusions, because (to mention one of several reasons) any point drawn may represent several measurement values. The figures do not show a significant change of the risk factor values over time or over age except for smoking (cigarette years). The figures show that most risk factors are distributed slightly skewed to the right. The figures of the risk factor values plotted against age show a slight decrease of the variation. It seems that extreme risk factor values disappear over time.

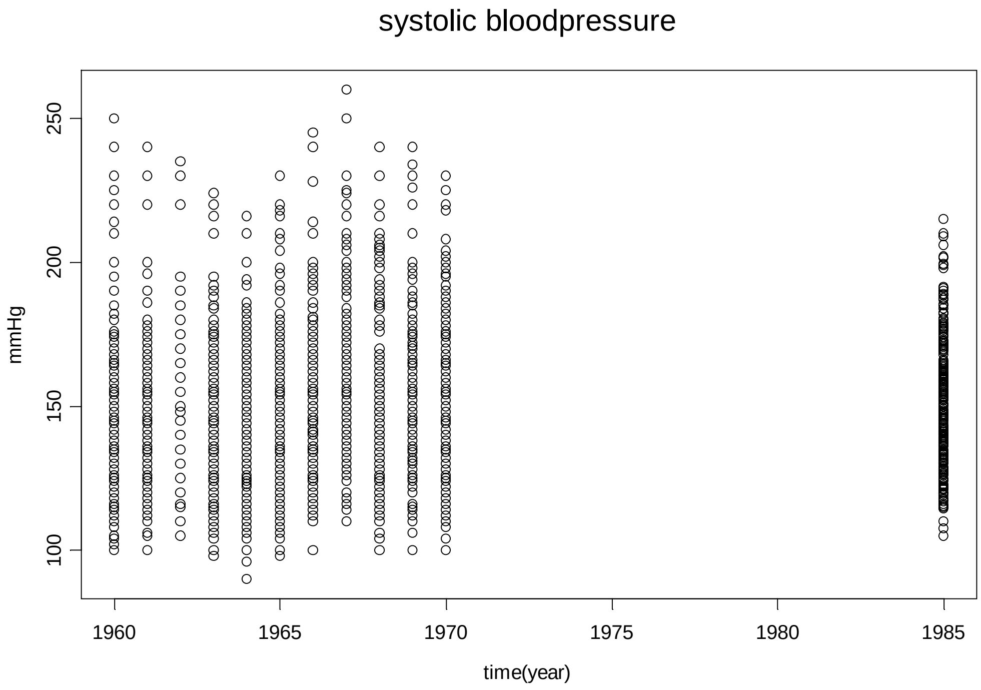


*Figure 1 Scatter plot of systolic bloodpresssure (mmHg) over time (year)*

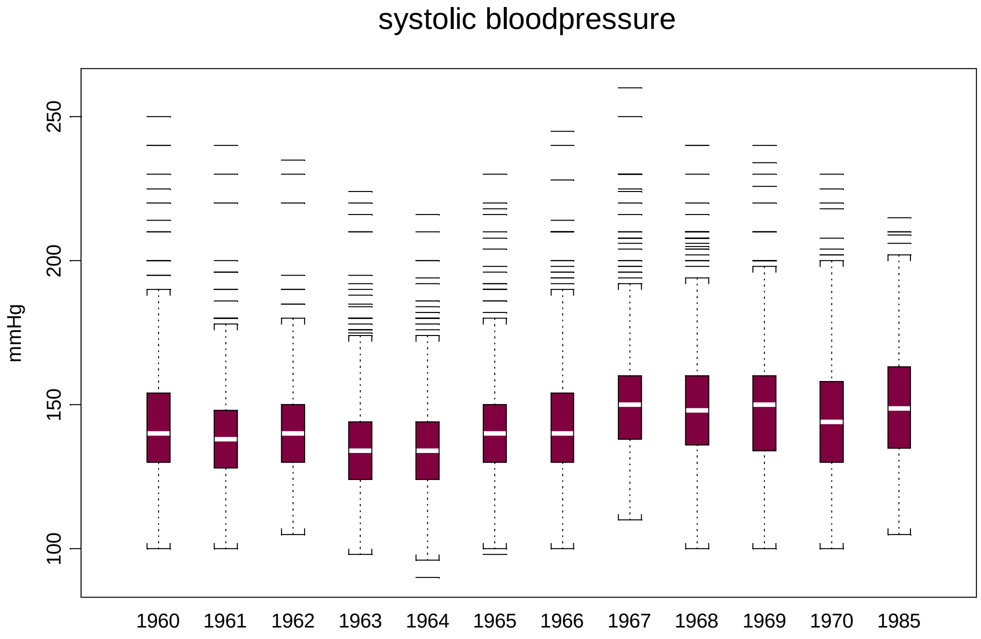


*Figure 2 Box plot of systolic bloodpresssure (mmHg) over time (year)*

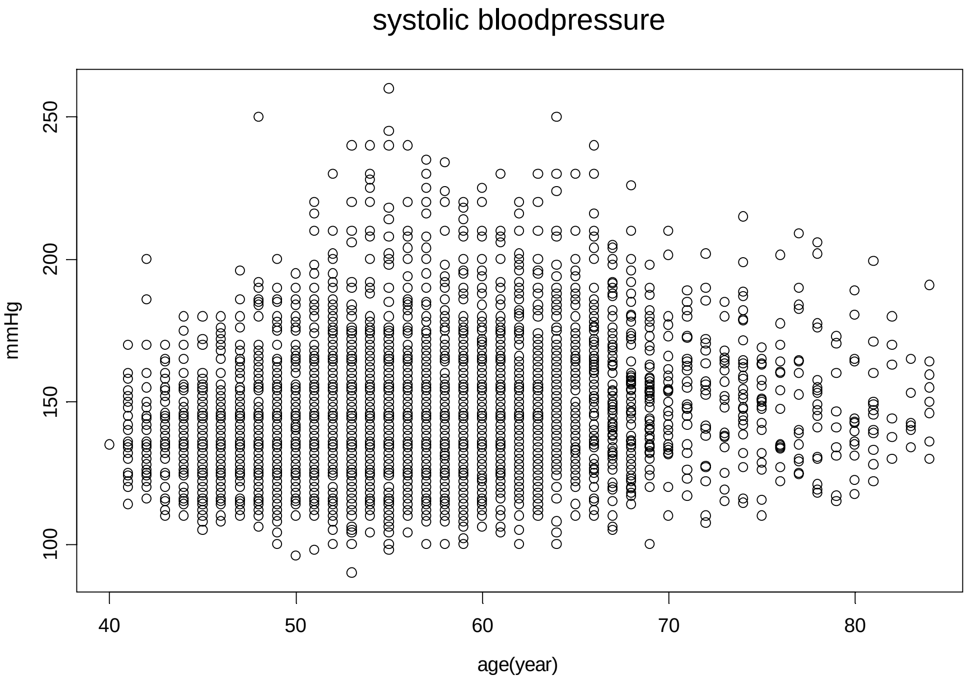


*Figure 3 Scatter plot of systolic bloodpressure (mmHg) over age (year)*

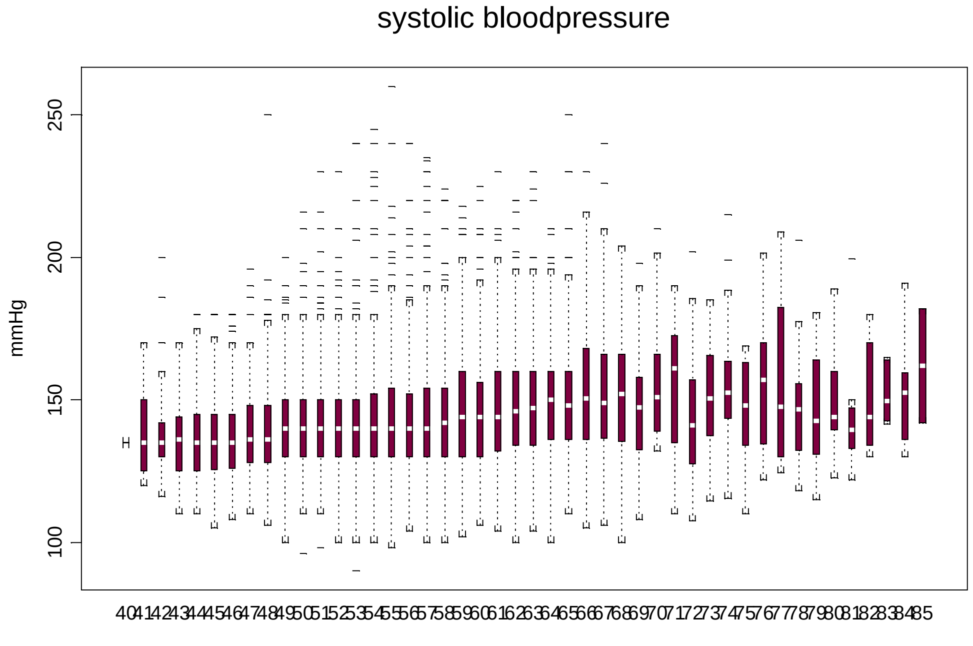


*Figure 4 Box plot of systolic bloodpresssure (mmHg) over age (year)*

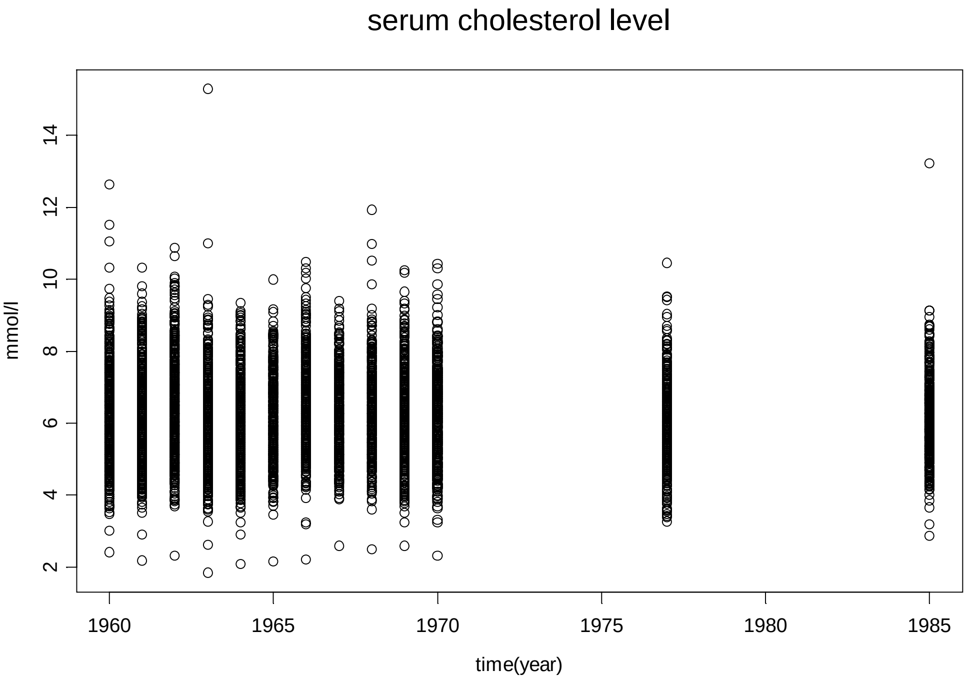


*Figure 5 Scatter plot of change in systolic bloodpressure (mmHg) over time (year)*

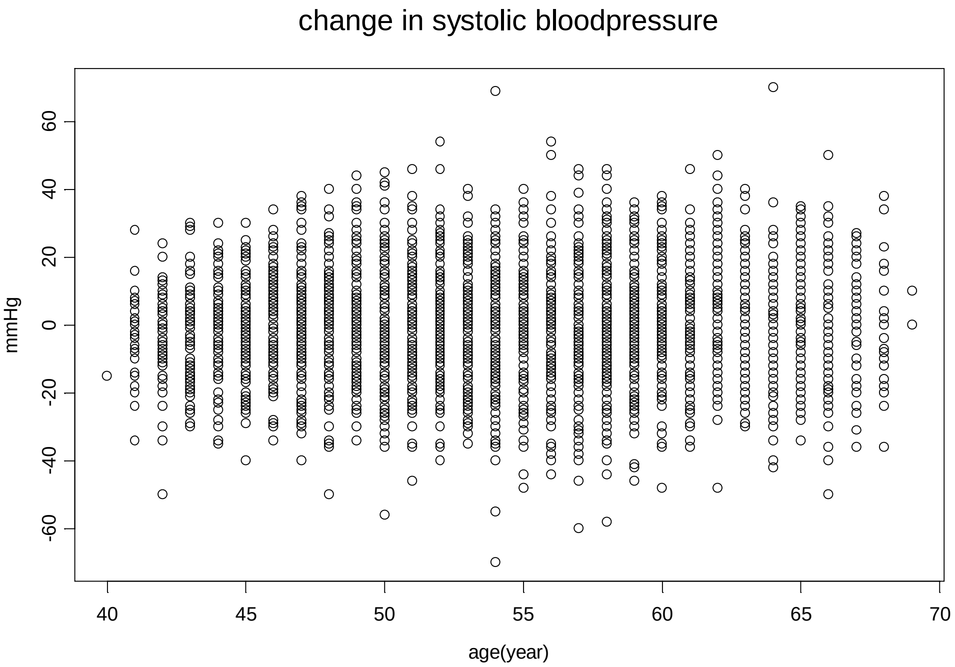


*Figure 6 Scatter plot of change in systolic bloodpressure (mmHg) over age (year)*

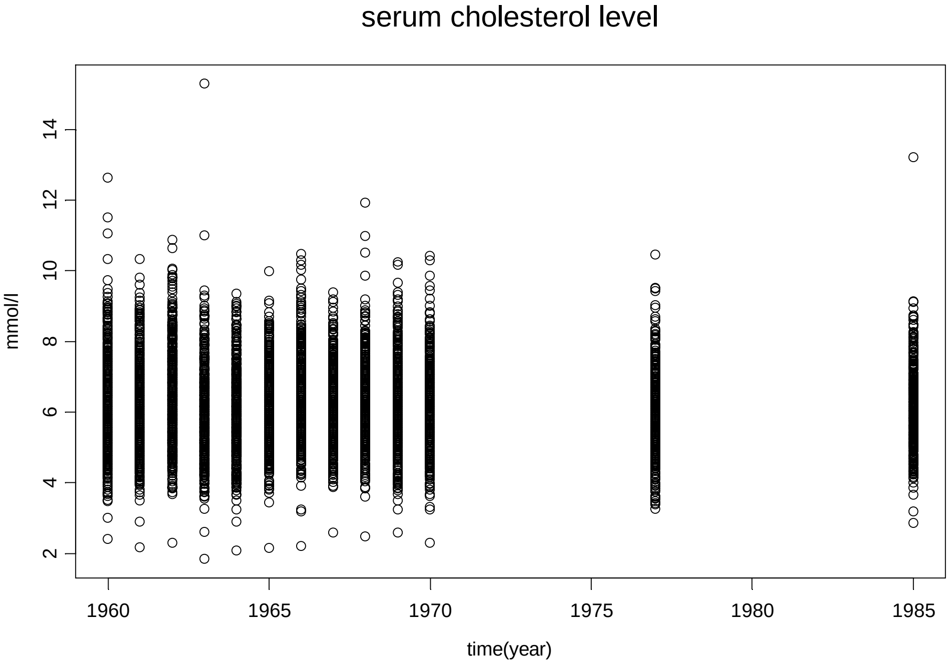


*Figure 7 Scatter plot of serum cholesterol level (mmol/l) over time (year)*

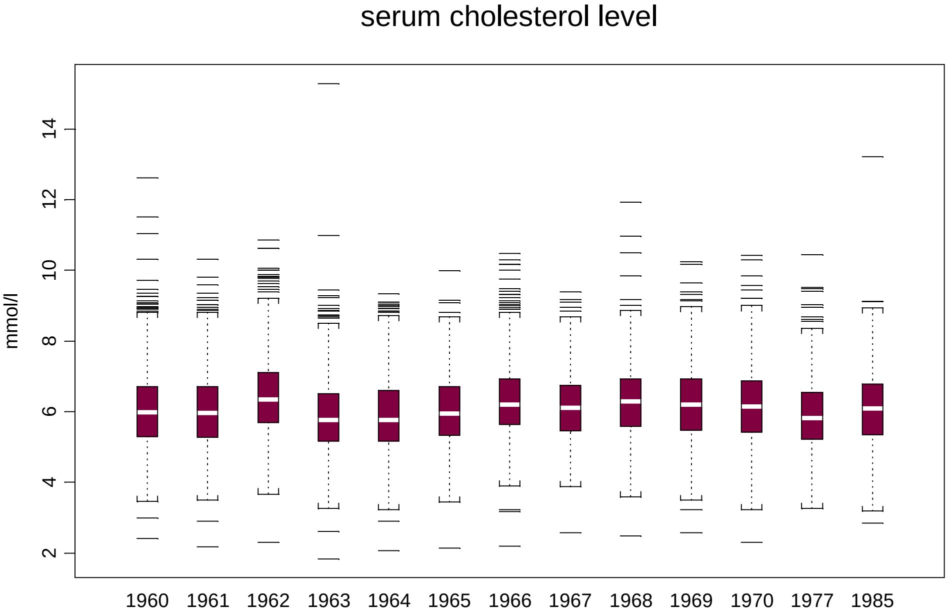


*Figure 8 Box plot of serum cholesterol level (mmol/l) over time (year)*

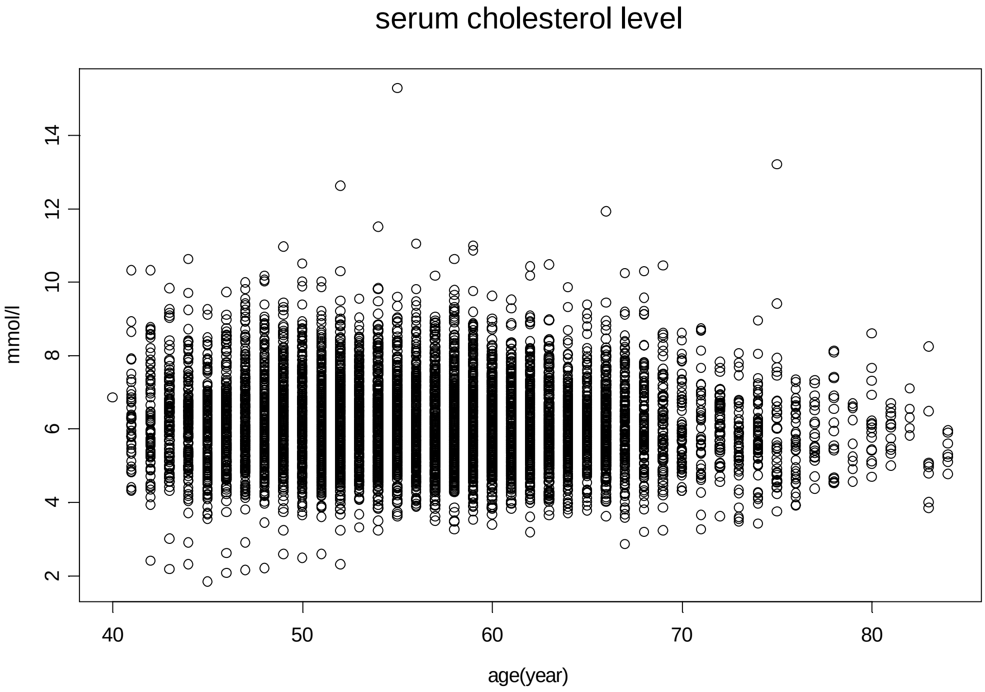


*Figure 9 Scatter plot of serum cholesterol level (mmol/l) over age (year)*

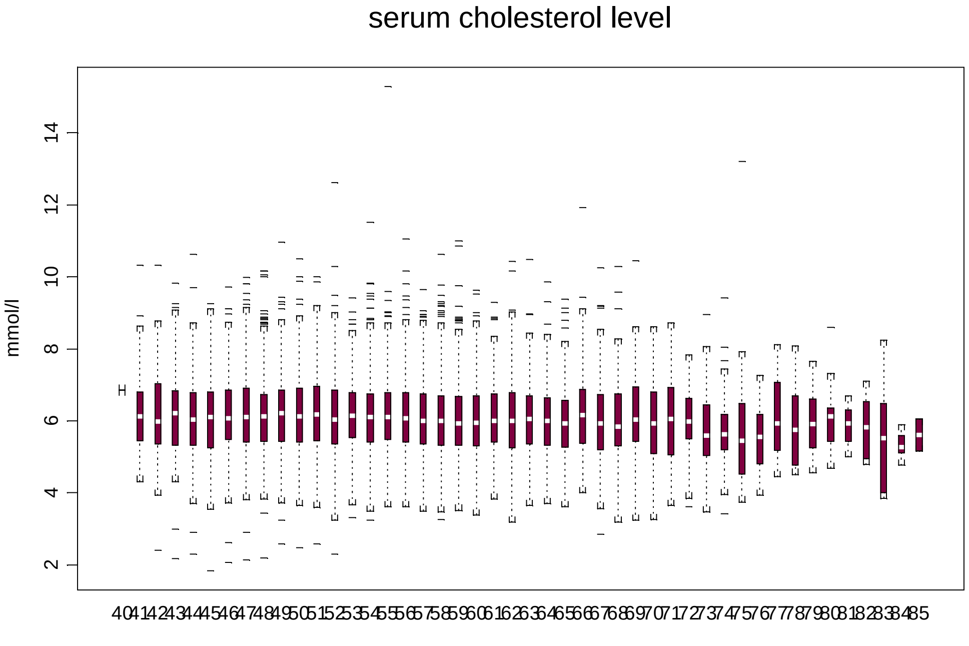


*Figure 10 Box plot of serum cholesterol level (mmol/l) over age (year)*

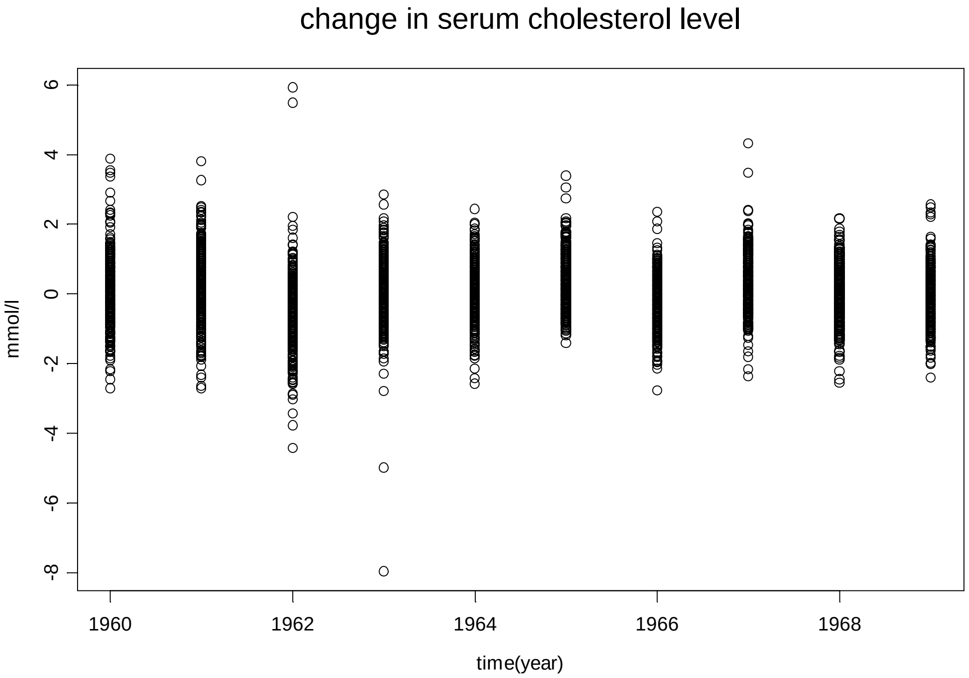


*Figure 11 Change in serum cholesterol level (mmol/l) over time (year)*

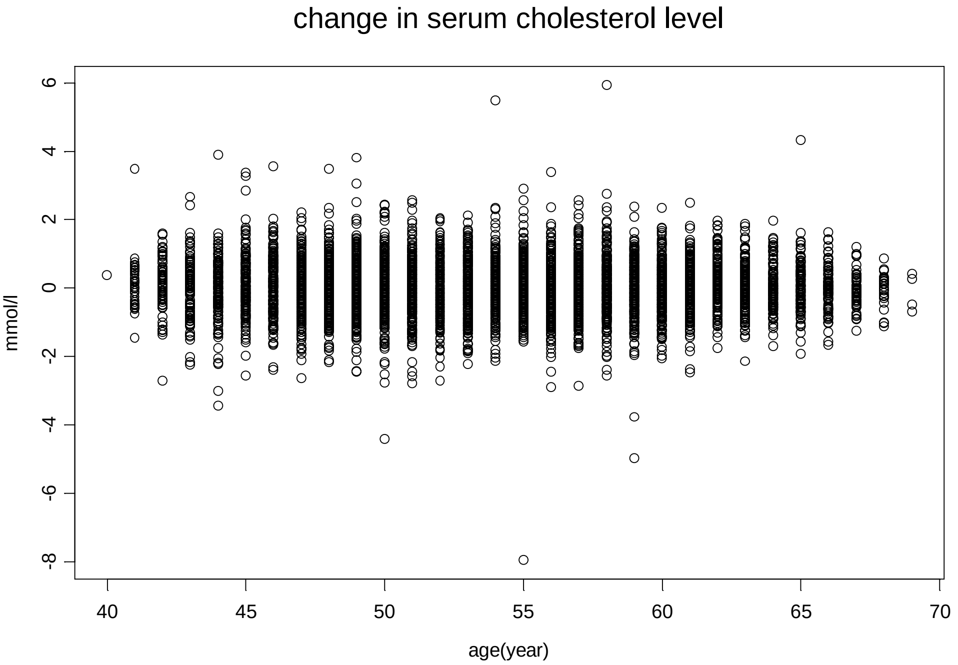


*Figure 12 Change in serum cholesterol level (mmol/l) over age (year)*

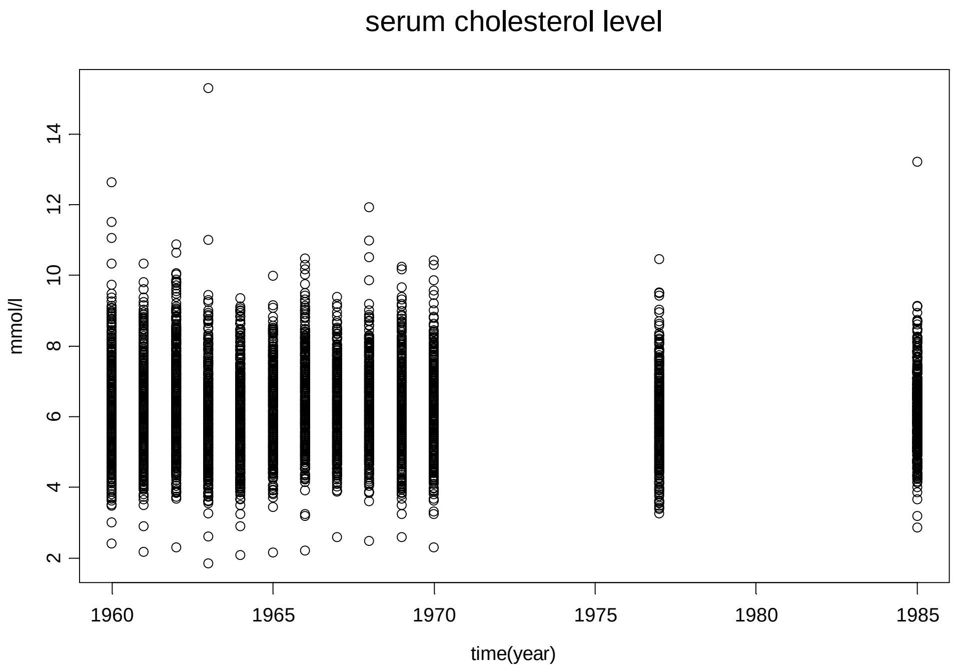


*Figure 13 Scatter plot of Body Mass Index (kg m$^{-2}$) over time (year)*

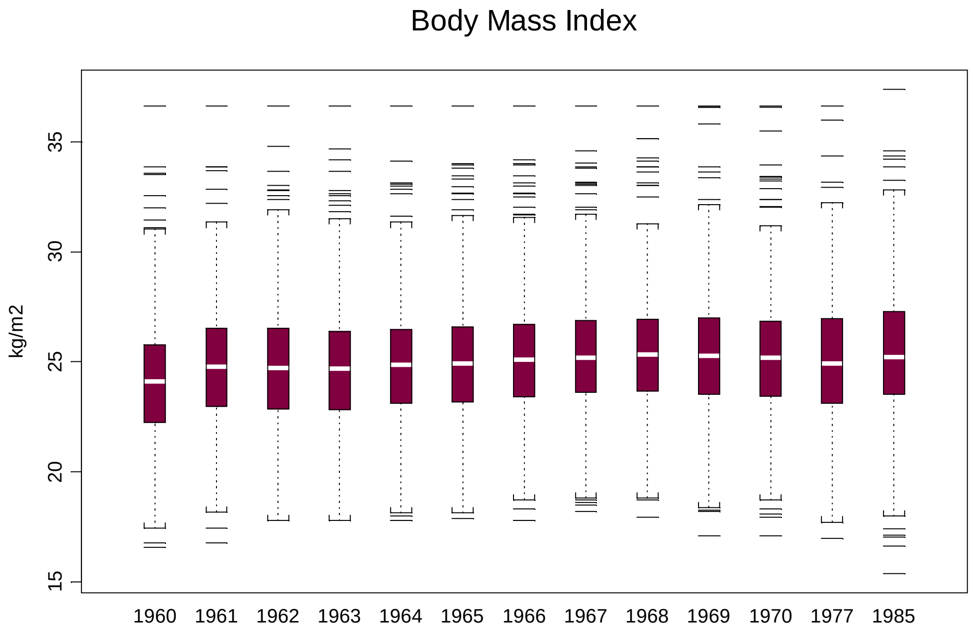


*Figure 14 Box plot of Body Mass Index (kg m$^{-2}$) over time (year)*

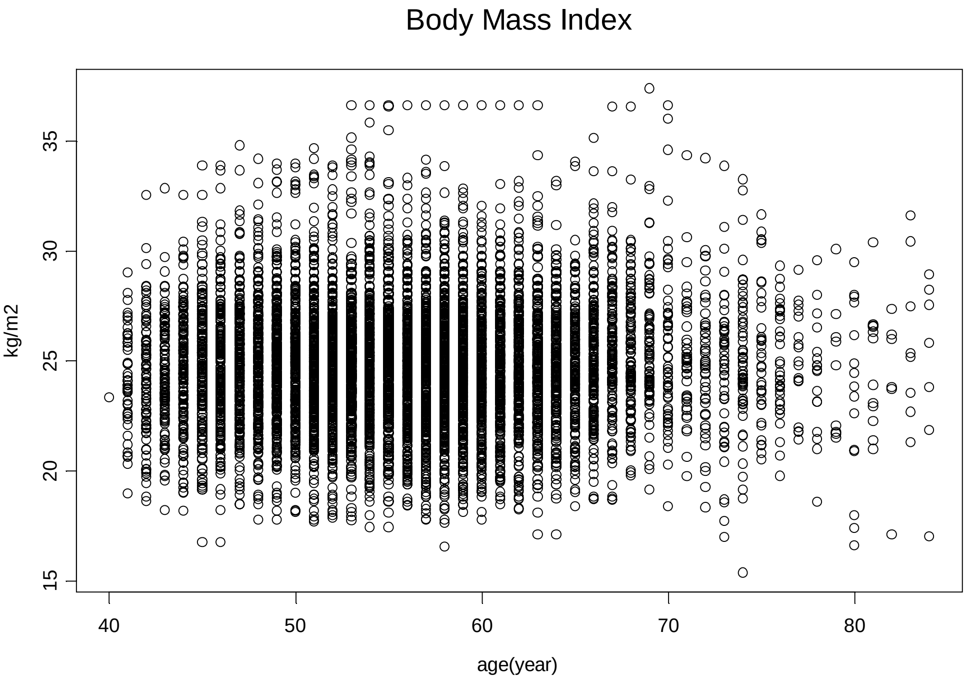


*Figure 15 Scatter plot of Body Mass Index (kg m$^{-2}$) over age (year)*

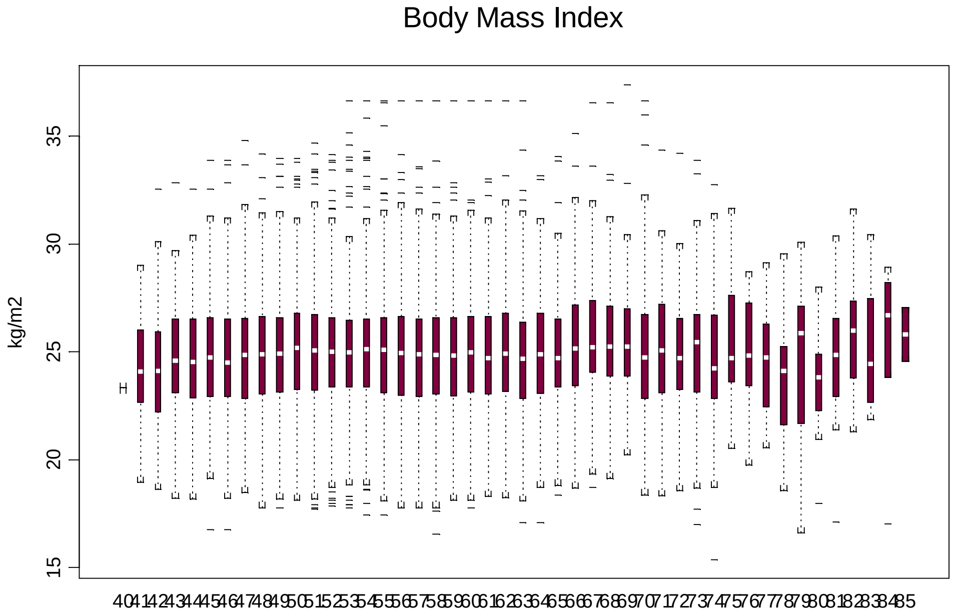


*Figure 16 Box plot of Body Mass Index (kg m$^{-2}$) over age (year)*

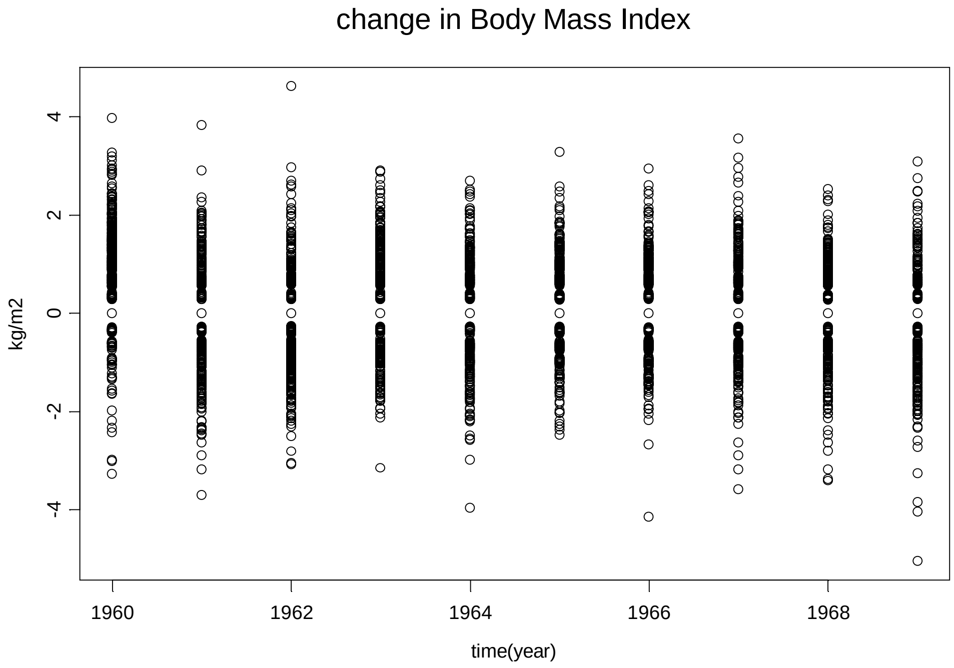


*Figure 17 Change in Body Mass Index (kg m$^{-2}$) over time (year)*

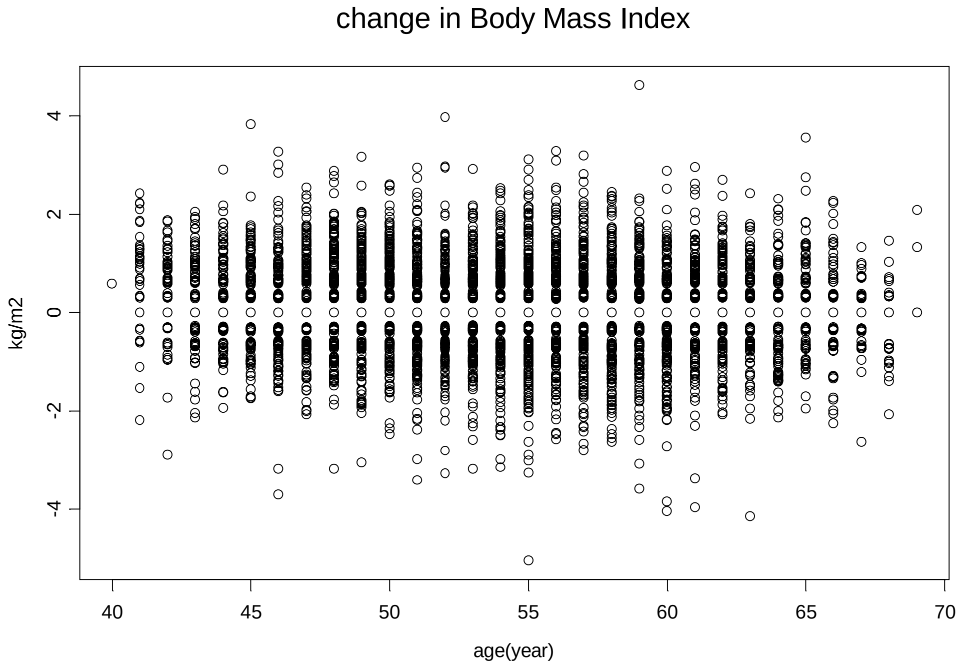


*Figure 18 Change in Body Mass Index (kg m$^{-2}$) over age (year)*

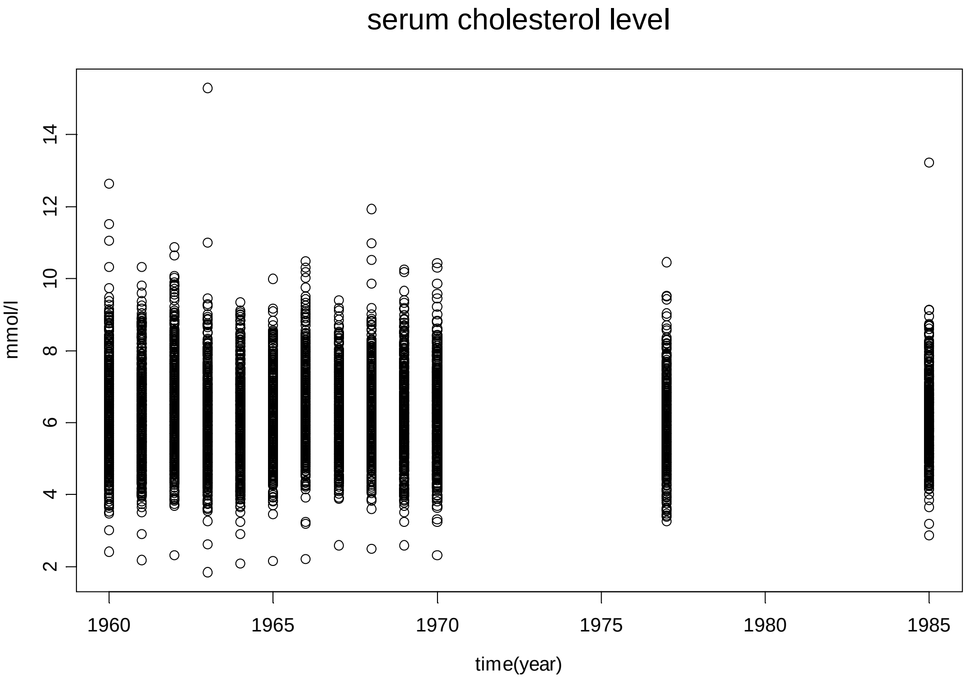


*Figure 19 Scatter plot of smoking (# cigarette years) over time (year)*

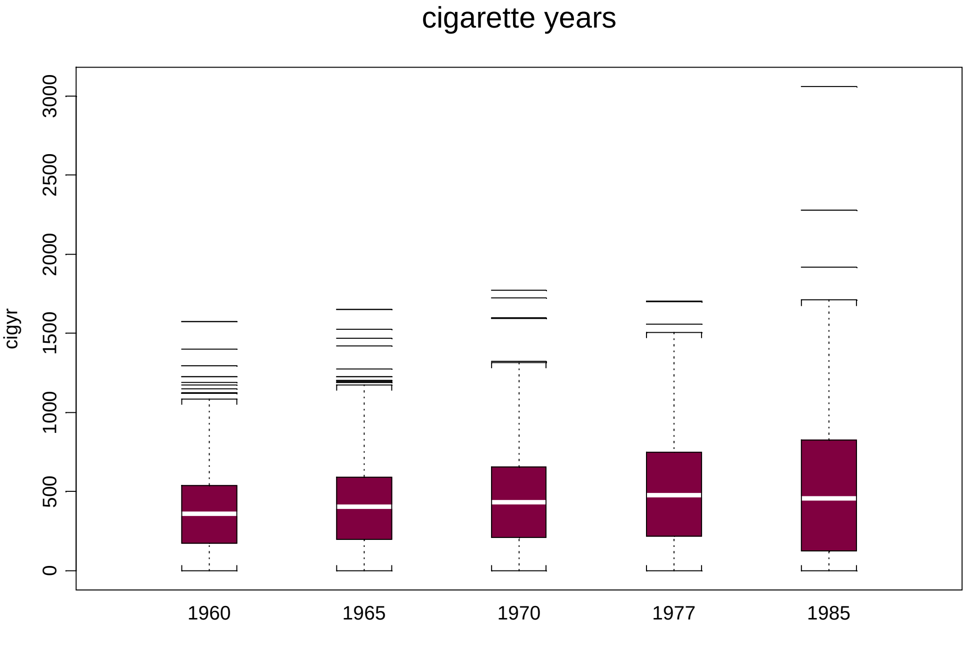


*Figure 20 Box plot of smoking (# cigarette years) over time (year)*

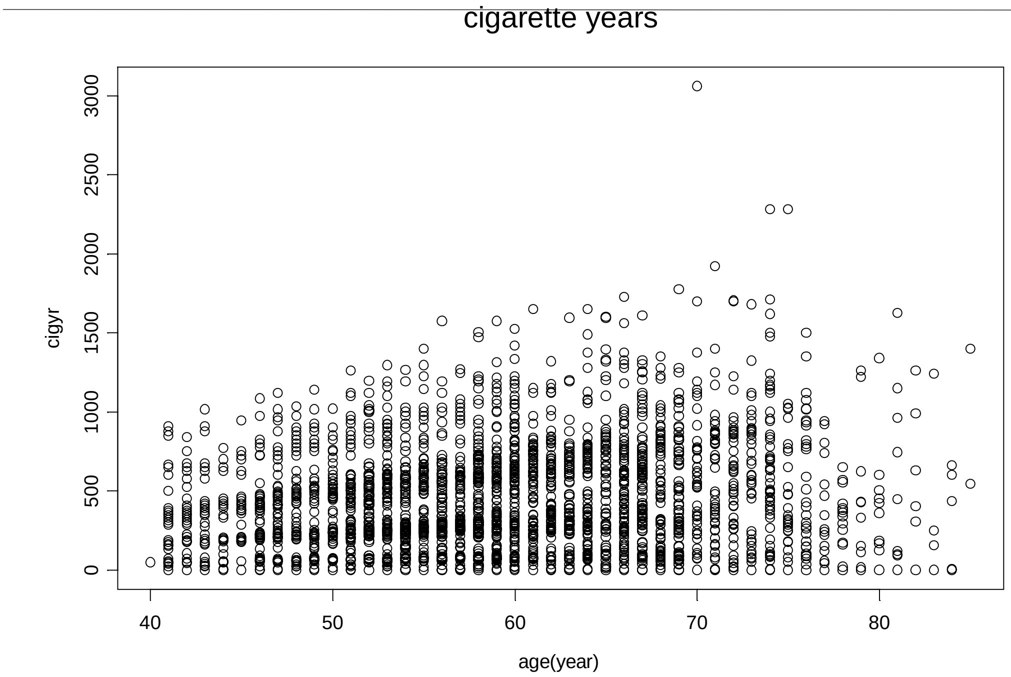


*Figure 21 Scatter plot of smoking (# cigarette years) over age (year)*

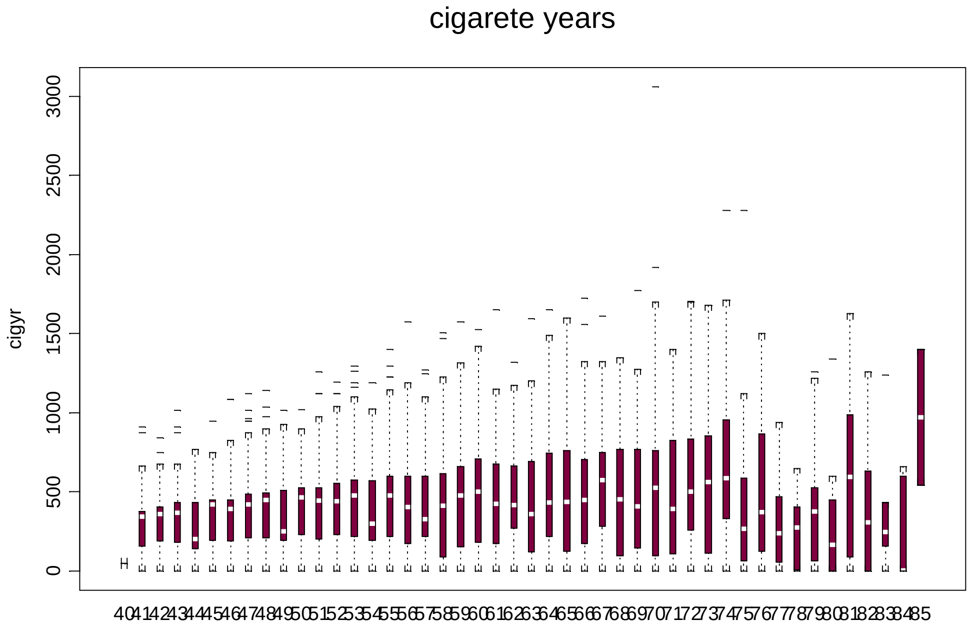


*Figure 22 Box plot of smoking (# cigarette years) over age (year)*

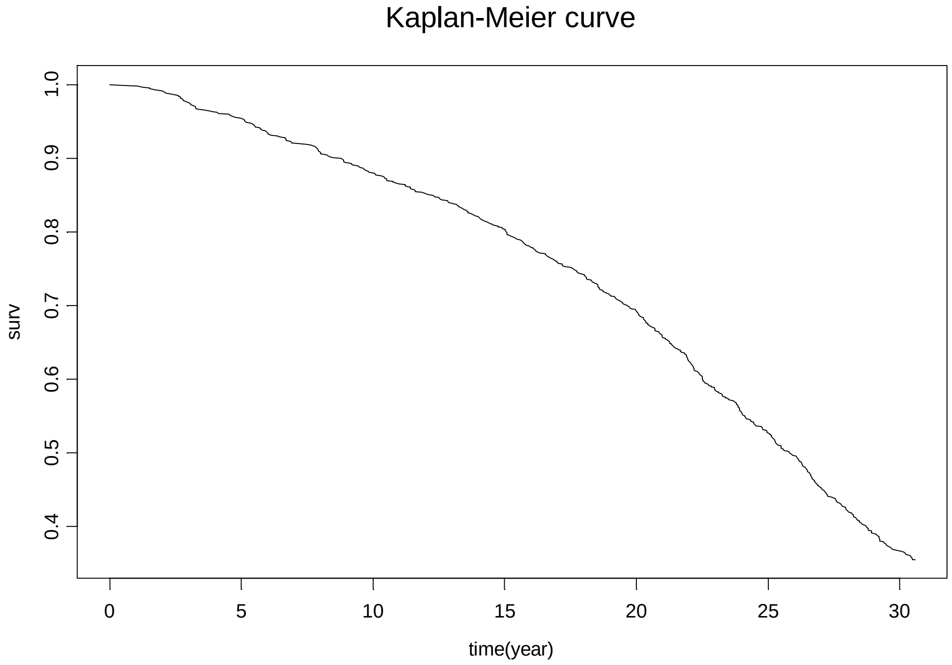


*Figure 23 Kaplan-Meier estimation of survival function over time*

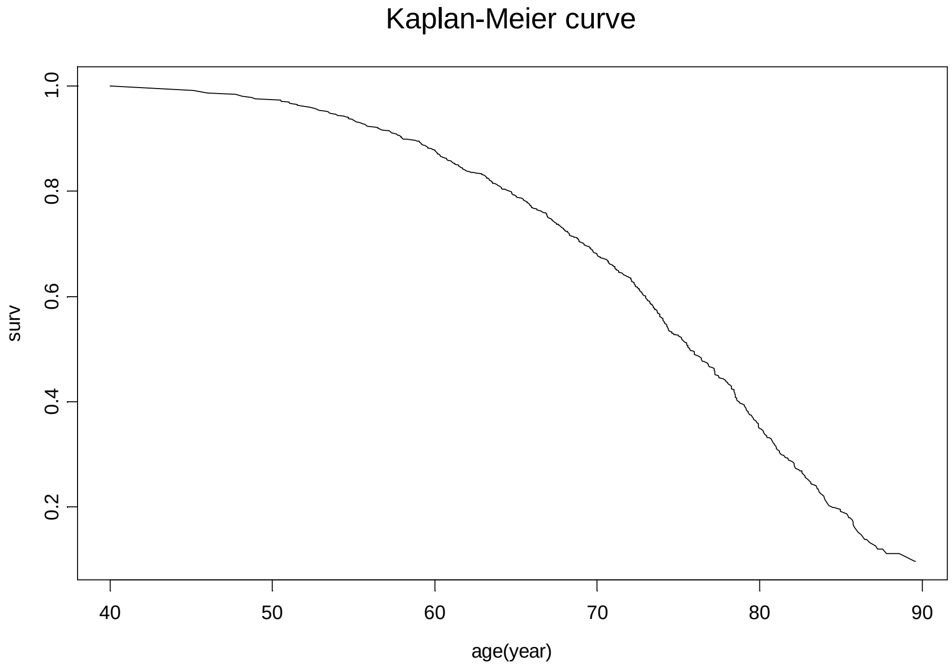


*Figure 24 Kaplan-Meier estimation of survival function over age*

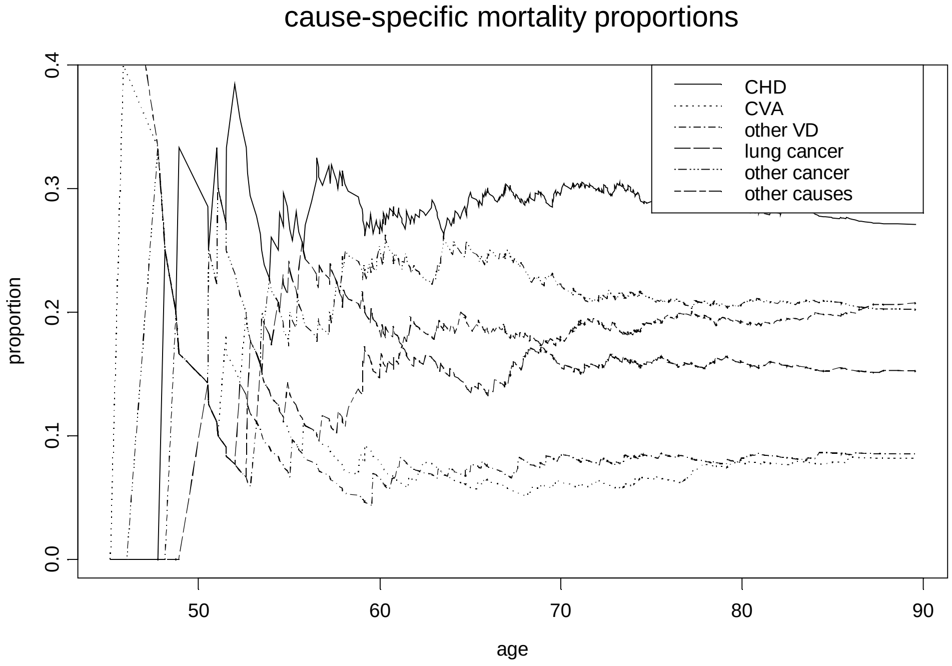


*Figure 25 Cause-specific mortality proportions over age*

Note: causes of death are in increasing order for the last age value recorded: cerebrovascular attack (CVA), other heart diseases, lung cancer, other cancers, other causes, and coronary heart diseases (CHD).

## 3.3 The causes of death distinguished

During follow-up measurements on mortality and morbidity have been recorded in the Zutphen-Study. In our analyses we have only used data on the outcome mortality. The variables having been used are:

- Year, month, and day of death. We have aggregated these data into one time of death, being the sum of year, (month-1)/12 and (day-1)/365.
- Cause of death, filled in according to the ICD (International Classification of Diseases) code (8th revision, 1965). The ICD codes have been aggregated into several mortality categories: coronary heart diseases (CHD): ICD 410-414; cerebrovascular attack (CVA, stroke): ICD 430-438; other heart diseases: rest numbers from ICD 390-459; lung cancer: ICD 162; other cancer types: rest numbers from ICD 140-208; other death risks.

In Figure 23 the Kaplan-Meier estimated survival function has been presented defined over time, in Figure 24 the one defined over age. The two survival functions are different due to population heterogeneity. In 1985 almost 50% of the cohort has died, in 1990 almost 65%.

*Table 6 The mortality numbers and proportions*

|  | **CHD** | **CVA** | **other heartd** | **lung-cancer** | **other cancer** | **other causes** | **total** |
|---|---|---|---|---|---|---|---|
| **until 1985** | 132<br>(31%) | 31<br>(7%) | 35<br>(8%) | 63<br>(15%) | 93<br>(22%) | 76<br>(18%) | 430<br>(100%) |
| **until 1990** | 153<br>(27%) | 46<br>(8%) | 48<br>(9%) | 86<br>(15%) | 114<br>(20%) | 117<br>(21%) | 564<br>(100%) |

Note: numbers between brackets are proportions; heartd: heart diseases.

In Table 6 the cause-specific mortality numbers and proportions until 1985 and 1990 respectively have been presented. In Figure 25 the proportions have been shown graphically as a function of age. Until age 60 years the cause-specific mortality proportions are very unstable due to small numbers. After age 60 years the proportions fluctuate around a trend or constant value. The CHD and cancer mortality proportions decrease and the other causes mortality proportion increases. The almost constant mortality proportions for older ages support the proportionality assumptions underlying the hazard functions of both the Cox model and the Manton&Stallard model.

We finish this paragraph with some overall conclusions. The Box plots over time and age suggest a systematic change until higher ages. These trends are also suggested by the correlation coefficients found, although the picture is not very clear. However, both Box plots and correlation coefficients may be misleading due to missing values (right censoring). The risk factors are distributed slightly skewed to the right, especially systolic bloodpressure and serum cholesterol level. However, we did not find a significant deviation from a normal distribution.

# 4. Results

## 4.1 Introduction

Before presenting the results of fitting the model of Manton&Stallard to the Zutphen-Study data we summarise the main model characteristics. The model describes the change of population risk factor levels over time, as the result of individual risk factor level changes and mortality. The mathematical equation that describes this change cannot be solved directly, due to its non-parametric form. Therefore we parameterise the model. We assume an initial multivariate normal distribution, a linear deterministic change within individuals, and a quadratic mortality hazard function. Then the risk factors are multivariately normally distributed over time. The baseline mortality hazard function describes the autonomous (exponential) increase of the hazard function over age

The model has been fitted by the method maximum likelihood. The likelihood function has been separated in two parts, one on the risk factor changes, assuming no mortality, and one on mortality, assuming no risk factor changes. Both model parts can be fitted separately. The submodel on mortality has been fitted in a two-stage procedure. First the optimal parameter of the autonomous change of the hazard function over age is estimated. Then, conditional on this parameter, the mortality submodel is fitted resulting in estimated values for the regression parameters of the hazard function.

For the 866-th individual no measurement values have been recorded, so it has been omitted from all analyses. For the 544-th, 669-th and 748-th individual no cause of death has been recorded, so they have been omitted from the analyses on mortality. Because the measurement time-intervals after 1970 are very long (until 1977 or 1985), we have used only the one-year time-intervals between 1960 and 1970 to fit the submodel on the risk factor changes. We have analysed how the results would change when including data for year 1977 with imputed values for the missing bloodpressure levels. In case of fitting the mortality submodel we have selected those subject-intervals with non-missing risk factor values at the start. Because most mortality events have taken place after 1970 we have included all subject-intervals after 1970. Because for 1977 no bloodpressure levels have been measured, we have treated 1970-1985 here as one time-interval.

## 4.2 Results on the risk factor level changes

The parameters that describe the risk factor changes have been estimated by the method of maximum likelihood. All data are grouped through the concept of subject-intervals. The change of the risk factors over each subject-interval is multivariately normally distributed conditional on the values at the start, i.e. $x_{ie}-x_{ib}|x_{ib} \propto N(A_0+A_1x_{ib},\Sigma)$, with: $x_{ib}$, $x_{ie}$: risk factor values at the start and end of the subject-interval respectively; $A_0$: the vector of constant risk factor changes; $A_1$: the matrix of changes proportional to the absolute values; $\Sigma = D^T D$: the variance-covariance matrix. Several model parameterisations have been analysed. These parameterisations differ in the configuration of the matrices $A_0$ and $A_1$, that describe the deterministic changes, and the matrix D, that describes the random changes. We have assumed that only $A_0$ is non-zero (assuming only constant deterministic changes), the diagonal of matrix $A_1$ is also non-zero (assuming linear deterministic changes without interactions), or all values are non-zero (assuming linear deterministic changes with interactions). For the upper-triangular matrix D we have assumed that the diagonal is non-zero (assuming random changes without interactions), or all values are non-zero (assuming random changes with interactions).

The explanatory variables chosen are systolic bloodpressure (SBP), serum cholesterol level, and Body Mass Index (BMI). We have only made use of the data until the year 1970. First we have presented results for all observation units with no missing values. In §4.3 we have presented results including subject-intervals with missing values (see §2.8). All model results have been presented in tabular form, i.e. the loglikelihood value, and the estimated values of the elements of the matrices $A_0$, $A_1$ and D. The coefficients of the vector $A_0$ describe the constant changes of the risk factors given by the row-name. The coefficients of matrix $A_1$ describe the changes of the levels of the risk factors given by the row-name by a unit change of the levels of the risk factors given by the column-names. For every model parameterisation results have been presented for all ages, age <=55 years, and age>55 years. All parameter estimates have been presented together with the standard errors. The standard errors have been estimated by using the Hessian matrix (..). In case of estimating only constant changes (Table 7) or full matrix A (Table 10) also the changes relative to the mean risk factor levels have been presented. The mean values used are: 145 mmHg (SBP), 6.10 mmol/l (chol) and 25.0 $kg/m^2$ (BMI). In Table 7 these relative changes have to be read such as: the net yearly increase relative to the mean level. In Table 10 the relative changes have to be read such as: the net yearly increase relative to the mean level (row name) attributable to a specific risk factor level (column name).

*Table 7 Constant deterministic and random changes without interactions*

| | | | **constant changes** | **SBP** | **cholesterol** | **BMI** |
|---|---|---|---|---|---|---|
| **ages** | | | **(intercepts)** | **regression coefficients (A), non-zero elements (D) respect.** | | |
| **all** | **loglikelihood** | -24329.1 | | | | |
| (n=6603) | **matrix A** | **SBP** | 3.8 (1.7) E-1 (2.6‰) | | | |
| | | **chol** | 9.8 (9.1) E-3 (1.6‰) | | | |
| | | **BMI** | 1.0 (0.1) E-1 (4.0‰) | | | |
| | **matrix D** | **SBP** | | 1.4 (0.0) E1 | | |
| | | **chol** | | | 7.4 (0.1) E-1 | |
| | | **BMI** | | | | 8.7 (0.1) E-1 |
| **<=55** | **loglikelihood** | -14118.0 | | | | |
| (n=3833) | **matrix A** | **SBP** | 2.2 (2.2) E-1 (1.5‰) | | | |
| | | **chol** | 1.4 (1.2) E-2 (2.3‰) | | | |
| | | **BMI** | 1.4 (0.1) E-1 (5. ‰6) | | | |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 7.7 (0.1) E-1 | |
| | | **BMI** | | | | 8.6 (0.1) E-1 |
| **>55** | **loglikelihood** | -10175.1 | | | | |
| (n=2770) | **matrix A** | **SBP** | 5.9 (2.7) E-1 (4.1‰) | | | |
| | | **chol** | 3 (13) E-3 (4.9‰) | | | |
| | | **BMI** | 5.2 (1.7) E-2 (2.1‰) | | | |
| | **matrix D** | **SBP** | | 1.4 (0.0) E1 | | |
| | | **chol** | | | 7.0 (0.1) E-1 | |
| | | **BMI** | | | | 8.7 (0.1) E-1 |

Notes: SBP: systolic bloodpressure, chol: serum cholesterol level, BMI: Body Mass Index; numbers within brackets: standard errors, and (for intercepts) also changes relative to mean values.

*Table 8 Constant deterministic and random changes with interactions between random changes*

| | | | **constant changes** | **SBP** | **cholesterol** | **BMI** |
|---|---|---|---|---|---|---|
| **ages** | | | **(intercepts)** | **regression coefficients (A), non-zero elements (D) respect.** | | |
| **all** | **loglikelihood** | -23416.1 | | | | |
| | **matrix A** | **SBP** | 3.6 (0.1) E1 | -2.5 (0.1) E-1 | | |
| | | **chol** | 1.4 (0.0) | | -2.3 (0.1) E-1 | |
| | | **BMI** | 1.5 (0.1) | | | -5.5 (0.4) E-2 |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 7.0 (0.1) E-1 | |
| | | **BMI** | | | | 8.5 (0.1) E-1 |
| **<=55** | **loglikelihood** | -13520.5 | | | | |
| | **matrix A** | **SBP** | 4.1 (0.2) E1 | -2.9 (0.1) E-1 | | |
| | | **chol** | 1.5 (0.1) | | -2.4 (0.1) E-1 | |
| | | **BMI** | 1.5 (0.1) | | | -5.4 (0.5) E-2 |
| | **matrix D** | **SBP** | | 1.2 (0.0) E1 | | |
| | | **chol** | | | 7.2 (0.1) E-1 | |
| | | **BMI** | | | | 8.4 (0.1) E-1 |
| **>55** | **loglikelihood** | -9823.5 | | | | |
| | **matrix A** | **SBP** | 3.4 (0.2) E1 | -2.3 (0.1) E-1 | | |
| | | **chol** | 1.3 (0.1) | | -2.1 (0.1) E-1 | |
| | | **BMI** | 1.4 (0.1) | | | -5.6 (0.6) E-2 |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 6.6 (0.1) E-1 | |
| | | **BMI** | | | | 8.6 (0.1) E-1 |

Notes: SBP: systolic bloodpressure, chol: serum cholesterol level, BMI: Body Mass Index; numbers within brackets: standard errors

*Table 9 Linear deterministic changes without interactions between the risk factors and constant random changes without interactions*

| | | | **constant changes** | **SBP** | **cholesterol** | **BMI** |
|---|---|---|---|---|---|---|
| **ages** | | | **(intercepts)** | **regression coefficients (A), non-zero elements (D) respect.** | | |
| **all** | **loglikelihood** | -23213.3 | | | | |
| | **matrix A** | **SBP** | 3.6 (0.1) E1 | -2.5 (0.1) E-1 | | |
| | | **chol** | 1.4 (0.0) | | -2.2 (0.1) E-1 | |
| | | **BMI** | 1.6 (0.1) | | | -6.1 (0.4) E-2 |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | 7.4 (0.9) E-2 | 1.3 (0.1) E-1 |
| | | **chol** | | | 6.9 (0.1) E-1 | 1.4 (0.1) E-1 |
| | | **BMI** | | | | 8.3 (0.1) E-1 |
| **<=55** | **loglikelihood** | -13424.9 | | | | |
| | **matrix A** | **SBP** | 4.1 (0.2) E1 | -2.9 (0.1) E-1 | | |
| | | **chol** | 1.5 (0.1) | | -2.4 (0.1) E-1 | |
| | | **BMI** | 1.7 (0.1) | | | -6.2 (0.5) E-2 |
| | **matrix D** | **SBP** | | 1.2 (0.0) E1 | 7.3 (1.2) E-1 | 1.1 (0.1) E-1 |
| | | **chol** | | | 7.2 (0.1) E-1 | 1.3 (0.1) E-1 |
| | | **BMI** | | | | 8.3 (0.1) E-1 |
| **>55** | **loglikelihood** | -9706.6 | | | | |
| | **matrix A** | **SBP** | 3.4 (0.2) E1 | -2.3 (0.0) E-1 | | |
| | | **chol** | 1.2 (0.1) | | -2.0 (0.1) E-1 | |
| | | **BMI** | 1.6 (0.2) | | | -6.0 (0.6) E-2 |
| | **matrix D** | **SBP** | | 1.4 (0.0) E1 | 7.8 (1.2) E-1 | 1.6 (0.2) E-2 |
| | | **chol** | | | 6.6 (0.1) E-1 | 1.6 (0.2) E-2 |
| | | **BMI** | | | | 8.3 (0.1) E-1 |

Notes: SBP: systolic bloodpressure, chol: serum cholesterol level, BMI: Body Mass Index; numbers within brackets: standard errors

*Table 10 Linear deterministic changes with interactions and constant random changes without interactions*

| | | | **constant changes** | **SBP** | **cholesterol** | **BMI** |
|---|---|---|---|---|---|---|
| **ages** | | | **(intercepts)** | **regression coefficients (A), non-zero elements (D) respect.** | | |
| **all** | **loglikelihood** | -23356.2 | | | | |
| | **matrix A** | **SBP** | 2.5 (0.2) E1 | -2.8 (0.1) E-1 (28%) | -1 E-3 ns | 6.1 (0.6) E-1 (11%) |
| | | **chol** | 1.2 (0.1) | 1 E-5 ns | -2.3 (0.1) E-1 (23) | 7.6 (3.4) E-3 (3%) |
| | | **BMI** | 1.7 (0.1) | -1.7 (0.6) E-3 (1%) | -2.7 (1.0) E-2 (1%) | -4.9 (0.4) E-2 (5%) |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 7.0 (0.1) E-1 | |
| | | **BMI** | | | | 8.5 (0.1) E-1 |
| **<=55** | **loglikelihood** | -13463.4 | | | | |
| | **matrix A** | **SBP** | 2.7 (0.2) E1 | -3.3 (0.1) E-1 (33%) | -1.7 E-1 ns | 7.8 (0.8) E-1 (13%) |
| | | **chol** | 1.2 (0.1) | 2.1 E-4 ns | -2.5 (0.1) E-1 (25%) | 1.3 (0.5) E-2 (5%) |
| | | **BMI** | 1.7 (0.2) | -9.7 (8.5) E-4 | -2.6 (1.3) E-2 (1%) | -5.0 (0.6) E-2 (5%) |
| | **matrix D** | **SBP** | | 1.2 (0.0) E1 | | |
| | | **chol** | | | 7.2 (0.1) E-1 | |
| | | **BMI** | | | | 8.5 (0.1) E-1 |
| **>55** | **loglikelihood** | -9807.0 | | | | |
| | **matrix A** | **SBP** | 2.5 (0.3) E1 | -2.5 (0.1) E-1) (25%) | 5 E-3 ns | 4.8 (0.1) E-1 (8%) |
| | | **chol** | 1.2 (0.1) | 2.1 E-4 ns | -2.1 (0.1) E-1 (21%) | 1.9 E-3 ns |
| | | **BMI** | 1.7 (0.2) | -1.1 (0.8) E-3 | -3.4 (1.6) E-2 (1%) | -5.1 (0.6) E-2 (5) % |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 6.6 (0.1) E-1 | |
| | | **BMI** | | | | 8.6 (0.1) E-1 |

Notes: SBP: systolic bloodpressure, chol: serum cholesterol level, BMI: Body Mass Index; numbers within brackets: standard errors, and (for matrix $A_1$) the changes relative to the mean risk factor levels (%)

We have compared our results with some other figures on the change of the risk factor levels between ages 40 and 55 in the Netherlands (see Table 11). The model results have been calculated by multiplying the constant change (see Table 7) with the age-interval length (15 years). The ‘baseline’ values have been calculated by fitting a linear function to the empirical values in 1960. The ‘Monitoring Project’ values have been calculated by fitting a linear function to the reported mean values for the age classes 40-45 till 55-60.

*Table 11 Comparison of age-trends of risk factor levels*

| | **Zutphen-Study** | | **Monitoring Project** |
|---|---|---|---|
| | **model** | **baseline** | |
| **SBP** | 5.7 | 6.6 | 7.9 |
| **cholesterol** | .15 | <0 | .23 |
| **BMI** | 1.6 | <0 | 0.5 |

Notes: SBP: mmHg, cholesterol: mmol/l, BMI: kg/m$^2$, Monitoring Project: Monitoring Project on Cardiovascular Disease Risk Factors, years 1987-1991: Verschuren et al., 1994.

The most surprising result of the analyses is that we have found increasing levels over age for all risk factors, although non-significant for cholesterol. Simple regression analyses resulted in decreasing levels for cholesterol and BMI (see Table 5). The main explanation for these differences is probably that mortality results in missing high risk factor levels for higher ages. Other possible explanations may be that curves over age based on cross-sectional studies can differ from those based on longitudinal studies, or disturbances due to random changes.

The order of risk factors by increasing relative change over age is: serum cholesterol level, systolic bloodpressure (SBP) and Body Mass Index (BMI). The relative changes are not constant over age and neither is their order. For example, for ages > 55 years SBP and BMI have changed positions, meaning that bloodpressure increase over age is larger for higher ages, but BMI increase is smaller. The change of the risk factor level changes over age suggests to include interaction terms with age in the regression model.

The random changes over a one-year time interval are relatively large, compared to the deterministic change (drift). The ratios vary from approximately 10 (for BMI), 40 (SBP) to 80 (cholesterol). Contrary to the deterministic changes, the random changes are almost constant over age. The random changes are relatively smallest for BMI. This result agrees with ‘common sense’: BMI is more stable than bloodpressure and cholesterol level within individuals. Of course the deterministic changes become more important for increasing time lengths. The interpretation of the random changes is not very clear. Next to diffusion (random change over time) variability is also introduced by measurement errors and biological variability (see also chapter 5).

The negative diagonal elements of matrix $A_1$ show that the one-year changes are larger for small risk factor levels. These negative regression coefficients are relatively smallest for BMI. These results can also be seen in the figures that have been presented in chapter 3. The negative linear relation is almost constant over age. The coefficients can be used to calculate the turning point, i.e. the risk factor level at which the change alters from an increase to a decrease. These points are 144, 140 and 119 mmHg for bloodpressure, 6.2, 6.2 and 6.1 mmol/l for cholesterol, 26.8, 27.6 and 25.8 kg/m$^2$ for BMI, for all ages, ages≤55, and ages>55 year respectively. The change of SBP and cholesterol level also depends on the BMI level, especially for younger ages. BMI changes are almost independent on the other risk factors. Analogously to the random changes described before, the deterministic results could be disturbed by measurement errors and biological variability (see also chapter 5).

We have used only one time parameter, i.e. age. We have assumed that the age-effects are constant over the whole time-interval (1960 to 1970). The age-trend we found may be biased by time trends. One way to correct for time trends could be to introduce time as an independent explanatory variable. However, due to the relatively small time length (10 years) it is questionable whether significant time trends will be found.

All model extensions (compared to the base model of non-zero vector $A_0$ and diagonal matrix D) have resulted in significantly better model fits. Most non-diagonal matrix elements (describing the interactions between the variables) were significant. Introducing non-zero elements in matrix D (interactions between random changes) resulted in a much larger model fit increase than introducing non-zero elements in matrix $A_1$ (interactions between deterministic changes). The aspect of interaction can be illustrated with the total and residual variances of the risk factor values (see Table 12). The residuals have not changed after including non-diagonal elements in the matrices $A_1$ (deterministic changes) or D (random changes). This means, that the model improvements are exclusively found with respect to the covariance structures, not with respect to the estimated values. The almost constant residual variances can be explained by measurement errors and biological variability (see chapter 5).

*Table 12 Total and residual variances of the risk factors*

| | | **SBP** | **chol** | **BMI** |
|---|---|---|---|---|
| total variance | | 374 | 1.23 | 7.49 |
| residual variance | table 8 | 168 | 0.53 | 0.74 |
| | table 9 | 169 | 0.53 | 0.74 |
| | table 10 | 166 | 0.53 | 0.74 |

Notes: for model parameterisations, see table 8, 9&10 respectively

We have also calculated the model parameters for smokers and non-smokers separately, assuming no interactions (see Table 8). Because smoking status has not been measured between

1960 and 1970 except for 1965, we have created new data sets for smokers and non-smokers in the following way. We have assumed that the smoking status had not changed during each five-year period when the status at the start and end were identical. In other words, for any individual each five-year observation period with identical smoking status at the end and start has generated five one-year subject-intervals. In all other cases (different status or missing values) we exclude the data from the new analyses. The estimated model parameters were almost identical for smokers and non-smokers. Therefore we have found no statistical reason to stratify the analyses by smoking status.

## 4.3 Including missing values

For each individual for each measurement point a value has been imputed if only one risk factor value was missing, and for 1977 bloodpressure levels have been imputed if all other risk factors were non-missing. Values have been imputed using a linear regression model that has been fit on all data points with no missing values. In each regression model we have included all two-way interaction terms. The estimated regression parameter values together with the standard errors have been presented in Table 13.

*Table 13 The multivariate normal distributions of the risk factors*

| | bloodpressure | cholesterol | BMI |
|---|---|---|---|
| intercept | 56 (18) | -3.07 (1.17) | 2.25 (.32) |
| age | .33 (.27) | .051 (.018) | .191 (.036) |
| bloodpressure | | .035 (.007) | .127 (.015) |
| cholesterol level | -.11 ns | | 1.42 (.24) |
| BMI | 2.72 (.71) | .35 (.04) | |
| age*bloodpressure | | .000 ns | -.0010 (.0002) |
| age*cholesterol | .055 (.025) | | -.0087 (.0036) |
| age*BMI | -.0037 (.0099) | -.0022 (.0006) | |
| bloodpressure*cholesterol | | | -.0041 (.0014) |
| bloodpressure*BMI | | -.0011 (.0002) | |
| cholesterol*BMI | -.072 (.068) | | |
| residual standard error | 18.0 | 1.09 | 2.58 |

Note: BMI: Body Mass Index; standard errors between brackets; ns: large p-value

In Table 14 the estimated values of the parameters of the deterministic and random change have been presented using the extended data set. We have only presented results for the model parameterisation with diagonal matrices $A_1$ and D, assuming no interactions between the linear deterministic and random changes of the risk factor levels.

*Table 14 Matrix A constant changes and diagonal elements and D diagonal*

| | | | **constant changes** | **SBP** | **cholesterol** | **BMI** |
|---|---|---|---|---|---|---|
| **ages** | | | **(intercepts)** | **regression coefficients (A), non-zero elements (D) respect.** | | |
| **all** | **loglikelihood** | -27581.1 | | | | |
| | **matrix A** | **SBP** | 4.2 (0.1) E1 | -2.9 (0.1) E-1 | | |
| | | **chol** | 1.4 (0.0) | | -2.3 (0.1) E-1 | |
| | | **BMI** | 1.5 (0.1) | | | -5.8 (0.4) E-2 |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 7.0 (0.1) E-1 | |
| | | **BMI** | | | | 9.2 (0.1) E-1 |
| **<=55** | **loglikelihood** | -13506.8 | | | | |
| | **matrix A** | **SBP** | 4.4 (0.0) E1 | -3.1 (0.1) E-1 | | |
| | | **chol** | 1.5 (0.1) | | -2.4 (0.1) E-1 | |
| | | **BMI** | 1.8 (0.1) | | | -6.6 (0.6) E-2 |
| | **matrix D** | **SBP** | | 1.2 (0.0) E1 | | |
| | | **chol** | | | 7.2 (0.1) E-1 | |
| | | **BMI** | | | | 9.4 (0.1) E-1 |
| **>55** | **loglikelihood** | -13960.8 | | | | |
| | **matrix A** | **SBP** | 4.5 (0.3) E1 | -3.0 (0.2) E-1 | | |
| | | **chol** | 1.2 (0.1) | | -2.1 (0.1) E-1 | |
| | | **BMI** | 1.3 (0.1) | | | -5.3 (0.4) E-2 |
| | **matrix D** | **SBP** | | 1.3 (0.0) E1 | | |
| | | **chol** | | | 6.7 (0.1) E-1 | |
| | | **BMI** | | | | 9.0 (0.1) E-1 |

Notes: SBP: systolic bloodpressure, chol: serum cholesterol level, BMI: Body Mass Index; numbers within brackets: standard errors

The results have to be compared with those presented in Table 8 to see the effects of the data augmentation. The main differences are: much more data for age>55 years and relatively large changes for bloodpressure. The standard errors of all parameters have only slightly decreased. The differences between the results for ages≤55 years and age>55 years have become smaller, especially for bloodpressure. The one linear model that has been used to impute bloodpressure levels over all ages was probably too simple. We conclude that in our case augmenting the data set with mainly imputed data for one structurally missing variable, i.e. bloodpressure levels for year 1977, is not very meaningful and does more harm than good.

## 4.4 Proportional hazards analyses on mortality

Before presenting the results of fitting the mortality submodel of the Manton&Stallard model, we have shows some results of proportional hazards analyses. Because the mortality function of the Manton&Stallard model is fully parametric, we have also made the Cox proportional hazards model fully parametric using the same exponential baseline hazard function. We have analysed two model parameterisations, one assuming proportional cause-specific baseline hazard functions (using extra proportionality coefficients), and one without this assumption (see §2.4). The models have been fit by the method of maximum likelihood using the same concept of subject-intervals. Because the differences between the results of both model parameterisations were very small, we only present results for the case of different baseline hazard functions.

*Table 15 Regression coefficients of the proportional hazards model*

| | **Total** | **CHD** | **CVA** | **other heart diseases** | **lung cancer** | **other cancer** | **other causes** |
|---|---|---|---|---|---|---|---|
| **Baseline risk factor levels; different baseline hazard functions** | | | | | | | |
| loglikelih | -2600.9 | -2534.6 | -2381.7 | -2381.0 | -2452.1 | -2486.5 | -2498.6 |
| Age (E-1) | 0.97 | 0.86 | 1.20 | 1.13 | .89 | .85 | 1.13 |
| SBP (E-3) | **11 (2)** | **13 (4)** | **16 (7)** | 6.6 na | 6.9 (6.3) | 1.9 ns | **16 (5)** |
| chol (E-2) | 6.2 (3.9) | **20 (7)** | 17 (13) | 23 na | 0.5 ns | -2 ns | -9 (9) |
| BMI (E-2) | -1.4 (1.9) | 5.4 (3.4) | -3.9 (6.4) | 1.7 na | -5.2 (4.8) | -1.5 ns | **-9.3 (4.0)** |
| Smoking[1] | 2.7 (1.1) | **4.2 (2.1)** | 2.5 (3.6) | 0.0 na | **11 (4)** | -0.0 ns | 0.0 ns |
| prop | 4.0 E-5 | 2.7 E-6 | 4.6 E-7 | 6.4 E-7 | 6.9 E-5 | 2.9 E-4 | 4.2 E-5 |
| **Current risk factor levels; different baseline hazard functions** | | | | | | | |
| Loglikelih | -1774.3 | -1728.0 | -1634.6 | -1638.0 | -1660.8 | -1699.1 | -1687.4 |
| Age (E-1) | 1.05 | 0.96 | 1.24 | 1.01 | .94 | 1.00 | 1.27 |
| SBP (E-3) | **8.0 (2.6)** | 6.6 (4.8) | 19 na | -4.3 ns | -4.2 ns | 5.0 ns | **20 (6)** |
| chol (E-2) | **12 (5)** | **38 (9)** | 4.8 na | -8.4 ns | 3.8 ns | -12 (12) | 10 ns |
| BMI (E-2) | -3.8 (2.1) | 6.3 (3.6) | -3.9 na | 6.9 (6.5) | **-14 (6)** | -3.0 ns | **-20 (5)** |
| Smoking[1] | 1.2 (1.1) | 2.7 (2.1) | 1.5 na | 1.9 ns | **8.1 (3.5)** | -3.1 (2.4) | -1.6 ns |
| prop | 3.3 E-5 | 6.4 E-7 | 3.6 E-7 | 1.2 E-5 | 1.5 E-3 | 1.3 E-4 | 2.5 E-5 |

Notes: ns: large p-value; na: non-estimable because of singular Hessian matrix, points at relatively large variances and/or covariances; significant parameters are presented in bold

For several risk factors and causes of death consistent significant parameter estimates have been found in both models. These are bloodpressure for total mortality and mortality due to other causes, cholesterol level for CHD mortality, Body Mass Index for CHD mortality and mortality due to other causes, and smoking for lung cancer. When using current instead of baseline values bloodpressure becomes non-significant for CHD and CVA mortality, and smoking for total and CHD mortality. However, BMI becomes significant for lung cancer mortality. In the case of cholesterol and bloodpressure high levels may be biased downwards due to medication. In the case of BMI the relation may be reverse: decreasing BMI levels may point at latent tumours. We conclude that for most risk factors the estimated regression coefficient values decrease when using current instead of baseline measurement values.

The loglikelihood values when using baseline or current risk factor values were very different. The main reason is the difference in observation time periods. In case of using the baseline and current values all persons and subject-intervals respectively with non-missing values have been included in the likelihood function. For many persons for some risk factors for some time points risk factors were missing, and so the related subject-interval has been excluded from the likelihood function. Moreover, because smoking status has not been measured during the years 1961-1964 and 1966-1969, we have included only those subject-intervals for which the smoking status of the related person was the same in 1960 and 1965 or in 1965 and 1970 respectively. For these subject-intervals it has been assumed that the smoking status has not changed between 1960-1965 and 1965-1970 respectively.

## 4.5 Results on mortality

The ‘mortality submodel’ has been analysed for outcome total mortality, and for several specific causes of death. Because the model is fully parametric, we have included mortality for all other causes for each specific cause of death being analysed. The hazard ratio is a quadratic function of the explanatory variables. This function has been described by the matrix Q, that is uniquely defined by an upper-triangular matrix U: $Q_k = U_k^T U_k$, with: $U_k$: upper-triangular matrix; k: index with respect to causes of death. The matrix U has been filled row-wise. Initially only the first row of U has been filled, meaning there are no interactions included between the variables on the matrix U level. The columns of U represent the intercept value and the regression coefficients with respect to the systolic bloodpressure (SBP), serum cholesterol level, Body Mass Index (BMI), and smoking (yes/no). The risk factor data have been transformed using a Gompertz-type (exponential) baseline hazard function with slope coefficient ß (see §2.7). The transformation has been done using the mid-point for each time-interval. The model has been fit by the method of maximum likelihood. The coefficient ß of the baseline hazard function has been chosen such that the likelihood function is maximised. For each model parameterisation we have presented the following results:

| | |
|---|---|
| ß | optimal parameter of the Gompertz-type baseline hazard function |
| logl | the log-likelihood value |
| U | the elements of the upper triangular root of the matrix Q; only the first row of U is assumed non-zero |

logl and U have been given for the optimal parameter ß value. We have calculated results for two model parameterisations: one using the original quadratic hazard regression model, the other using a linear regression model to enable comparing the model results to those of the Cox model (see §2.4). Similarly to the Cox mortality analyses we have calculated results for several specific causes of death.

*Table 16 Regression coefficients using quadratic hazard regression model*

| | **U:** | | **intercept/** | **regression coefficients for** | | | |
|---|---|---|---|---|---|---|---|
| | | | **constant** | **SBP** | **cholesterol** | **BMI** | **smoking** |
| | **ß** | **loglikelihood** | **(E-2)** | **(E-5)** | **(E-3)** | **(E-3)** | **(E-3)** |
| total | .1003 | -1802.7 | 4.3 (2.2) | **28 (10)** | **3.7 (1.8)** | -1.3 (0.7) | 3.1 (4.0) |
| CHD | .1004 | -1758.2 | -7.0 (2.9) | 9.2 ns | **11 (3)** | 1.8 (1.0) | 5.8 (5.8) |
| CVA | .1010 | -1663.5 | -0.2 ns | **29 (13)** | -0.2 ns | -0.4 ns | 4.4 ns |
| other heart dis | .1002 | -1666.4 | 1.8 (3.5) | -11 ns | -1.4 ns | 1.5 (1.2) | 3.2 ns |
| lungcancer | .1004 | -1689.5 | 12 (4) | -10 ns | 0.3 ns | **-2.9 (1.2)** | **12 (6)** |
| other cancer | .0999 | -1728.4 | 6.4 (3.1) | 11 ns | -1.9 ns | 0.7 ns | -7.4 (5.8) |
| other causes | .1003 | -1714.6 | 9.3 (3.3) | **57 (15)** | 1.3 ns | **-5.4 (1.2)** | -9.4 (5.9) |

Notes: standard errors between brackets; ns: large p-value; significant parameters except for intercepts presented in bold

*Table 17 Regression coefficients using linear hazard regression model*

| | | | **intercept/** | **regression coefficients for** | | | |
|---|---|---|---|---|---|---|---|
| | | | **constant** | **SBP** | **cholesterol** | **BMI** | **smoking** |
| | **ß** | **loglikelihood** | **(E-3)** | **(E-5)** | **(E-5)** | **(E-5)** | **(E-4)** |
| total | .1003 | -1803.2 | 1.4 (3.2) | **4.0 (1.5)** | 49 (27) | -19 (11) | 3.6 ns |
| CHD | .1007 | -1760.0 | -8.5 (2.6) | 0.3 ns | **108 (25)** | **18 (9)** | 7.0 (6.2) |
| CVA | .1003 | -1663.9 | -0.7 ns | **1.9 (0.9)** | -6.6 ns | -3.1 ns | 3.7 ns |
| other heart dis | .1004 | -1666.8 | -0.6 ns | -0.9 (0.8) | -9.1 ns | 14 (9) | 2.6 ns |
| lungcancer | .1004 | -1690.1 | 8.8 (2.7) | -0.8 ns | -3.6 ns | **-24 (7)** | 4.0 (5.2) |
| other cancer | .1005 | -1729.6 | 3.3 (2.7) | 11 ns | -9.8 ns | -6.5 ns | -6.9 (5.7) |
| other causes | .1005 | -1709.3 | 9.6 (1.9) | **6.4 (1.2)** | 15 ns | **-6.7 (0.6)** | **-16 (4)** |

Notes: standard errors between brackets; ns: large p-value; significant parameters except for intercepts presented in bold

The results using the model of Manton&Stallard are not much different from those of the Cox proportional hazards model. The differences between the loglikelihood values are caused by the different ways that is dealt with autonomous age trends. In the Cox model age has been included as a covariate, in the model of Manton&Stallatd data have been used that have been transformed following a given age trend. The combinations of significant risk factors for specific causes of death are almost identical. The standard errors shown have been defined for the parameters of the matrix U. The standard errors of the corresponding linear and quadratic terms of the hazard ratios (matrix Q) can be approximated using the delta method (not shown here).

*Table 18 Regression coefficients using quadratic model with interaction terms*

| | **ß**<br>**loglikelihood** | | **constant** | **SBP** | **cholesterol** | **BMI** | **smoking** |
|---|---|---|---|---|---|---|---|
| **CHD** | .1004 | | (E-2) | (E-4) | (E-3) | (E-3) | (E-2) |
| | 1753.6 | U row 1 | -10.3 (4.5) | 1.0 ns | 8.3 (9.3) | 3.3 (4.1) | 0.4 ns |
| | | U row 2 | | -3.0 (2.6) | **-17 (4)** | **6.1 (1.6)** | **-2.8 (1.3)** |
| | | | (E-4) | (E-6) | (E-5) | (E-5) | (E-4) |
| | | Q row 1 | 107 | -11 | -87 | -35 | -4.2 |
| | | Q row 2 | -0.1 | 0.1 | 5.9 | -0.2 | 0.1 |
| | | Q row 3 | -8.7 | 5.9 | 35 | -7.4 | 5.1 |
| | | Q row 4 | -3.5 | -1.5 | -7.4 | 4.8 | -1.6 |
| | | Q row 5 | -4.2 | 8.9 | 51 | -16 | 8.1 |
| **other** | .0999 | | (E-2) | (E-4) | (E-3) | (E-3) | (E-2) |
| **causes** | 1711.5 | U row 1 | -8.0 (5.5) | -5.6 (3.6) | 0.1 ns | 4.7 (4.0) | 1.2 ns |
| | | U row 2 | | -3.5 (5.6) | -9.6 (7.0) | 5.0 (2.9) | -3.9 (2.9) |
| | | | (E-4) | (E-6) | (E-5) | (E-5) | (E-4) |
| | | Q row 1 | 63 | 44 | -0.9 | -38 | -9.4 |
| | | Q row 2 | 0.4 | 0.4 | 0.3 | -0.4 | 0.1 |
| | | Q row 3 | -0.1 | 3.3 | 9.2 | -4.7 | 3.8 |
| | | Q row 4 | -3.7 | -4.4 | -4.7 | 4.7 | -1.3 |
| | | Q row 5 | -9.5 | 7.2 | 3.8 | -1.4 | 17 |

Notes: standard errors between brackets; significant parameters presented in bold

We have also fitted the model with the first two rows of the upper part of matrix U being filled instead of only the first one for cause of death CHD and other causes respectively, thus including interactions on the matrix U level. The results, together with the resulting matrices Q, are presented in Table 18 and Figures 26 to 29.

When using a linear instead of quadratic hazard function and applying the formulas of §2.4, the regression parameters found agreed with those of the Cox proportional hazards model. However, hazard rate values became negative for some causes of death with small absolute risk values. When introducing interactions on the matrix U level for mortality due to CHD and other causes, the loglikelihood values increased only slightly. In case of mortality due to other causes, no regression coefficient is statistically significant in the quadratic model. The graphical results show that most relations are monotonic, with exception for BMI for low absolute risks. In case of CHD mortality the coefficient for the quadratic term for cholesterol is negative, resulting in U-shaped risk curves. This negative value was statistically significant. The interpretation of these U-shaped relations is not very simple. One has to take account of possible reverse relations: latent diseases may result in decreasing risk factor levels. The resulting hazard rate ratios are qualitatively similar for smokers and non-smokers, but we have found some differences. Especially for mortality due to other causes the risk differences are much larger for smokers than for non-smokers.

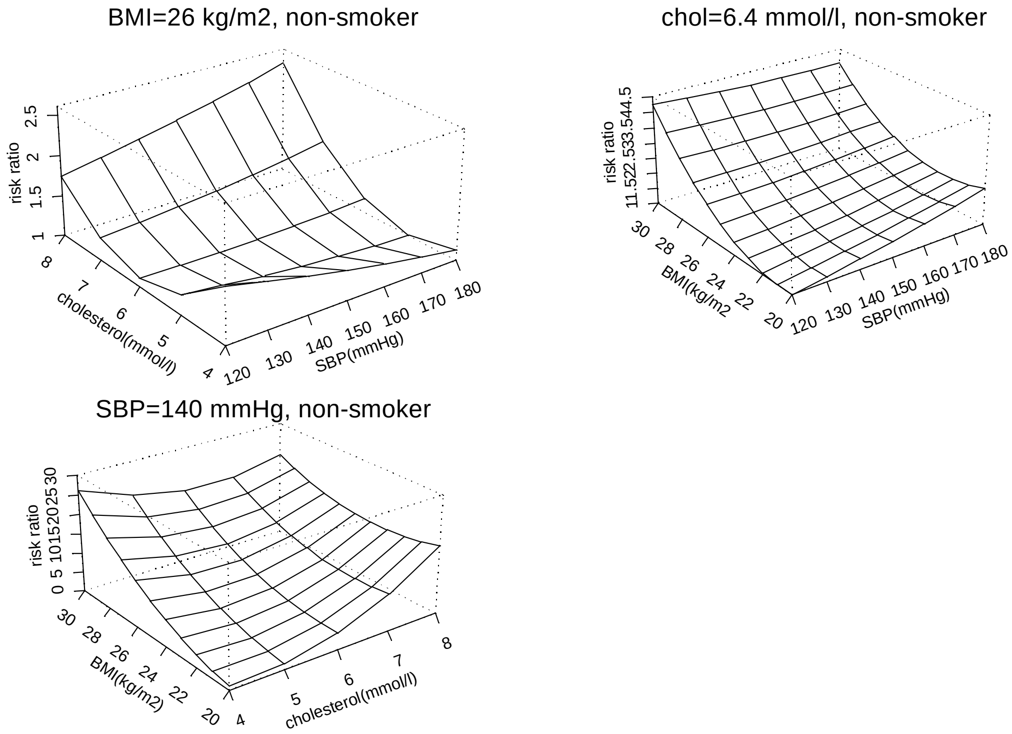


*Figure 26 Hazard rate ratios for mortality due to CHD for non-smokers*

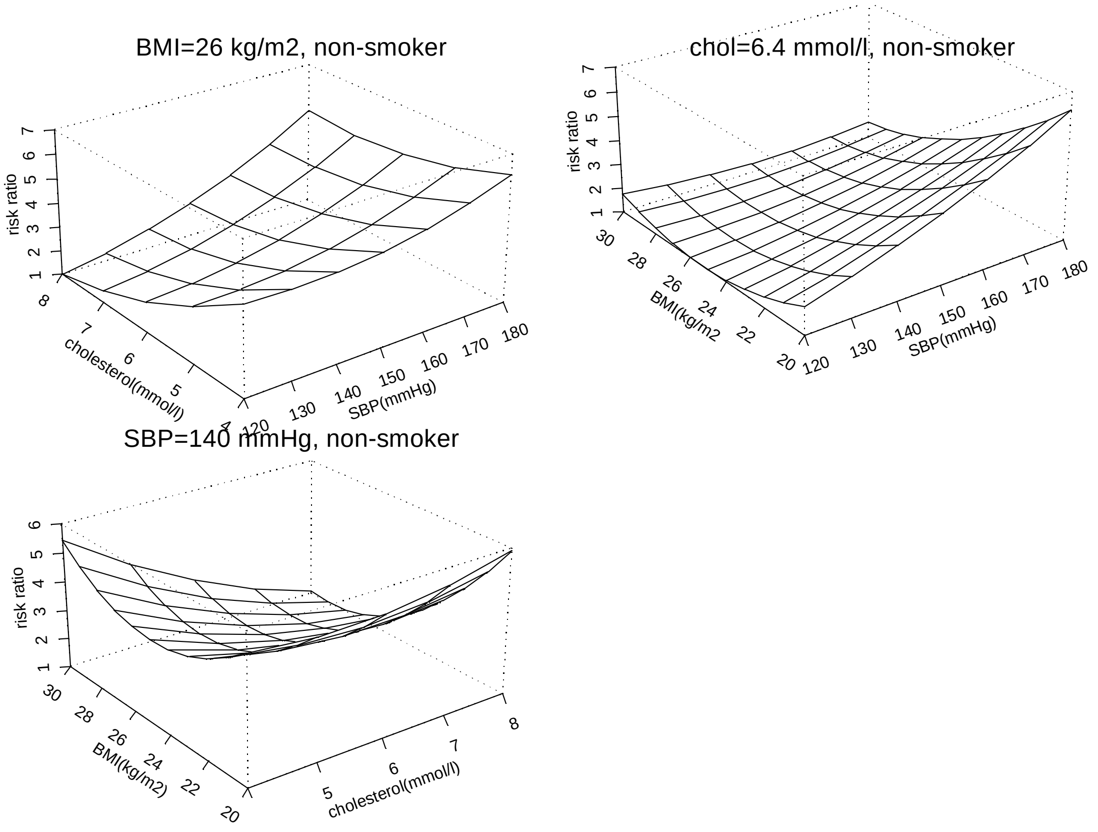


*Figure 27 Hazard rate ratios for mortality due to other causes for non-smokers*

Notes: all ratios defined relative with respect to smallest value within two-dimensional range.

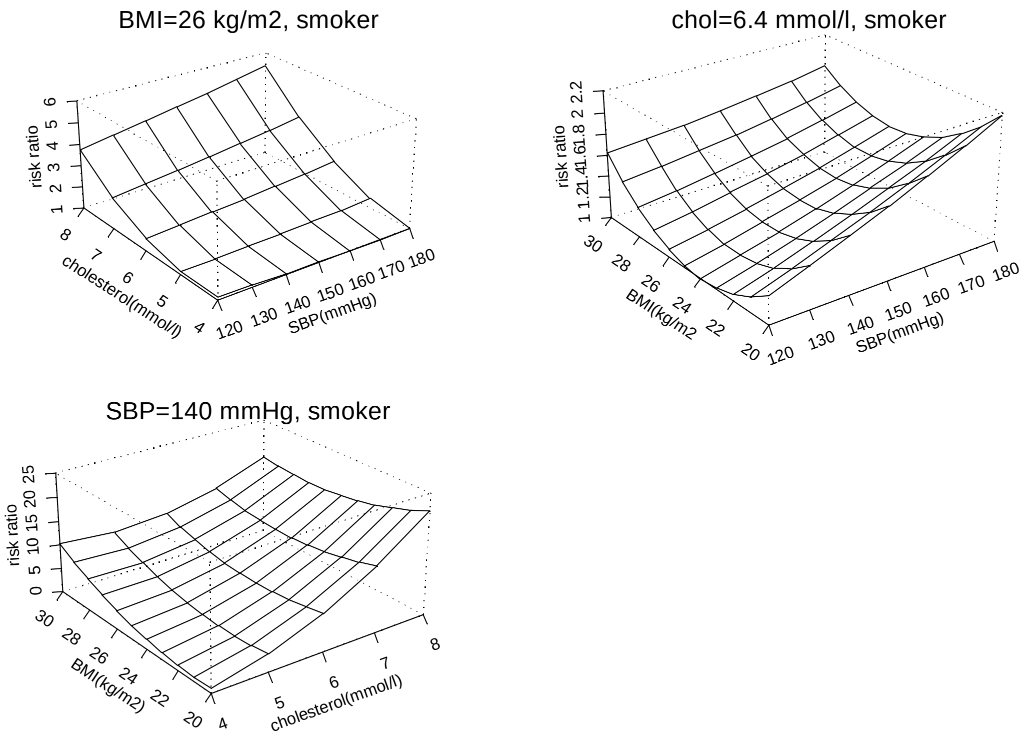


*Figure 28 Hazard rate ratios for mortality due to CHD for smokers*

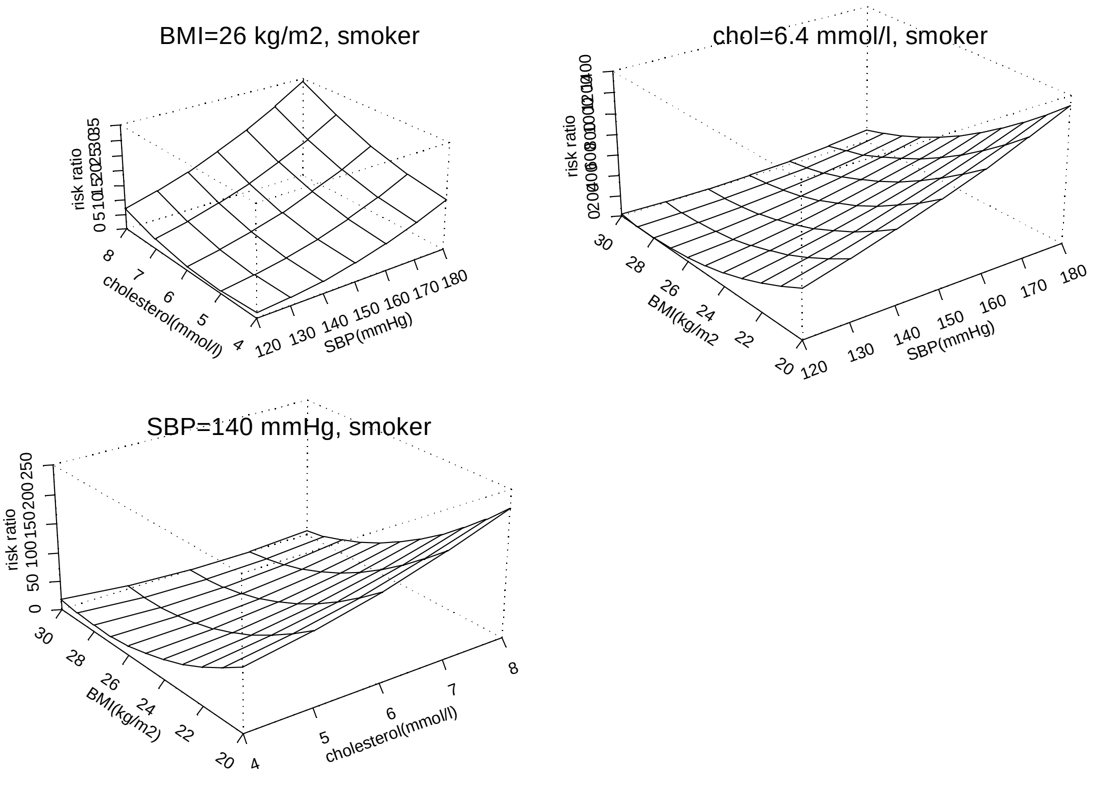


*Figure 29 Hazard rate ratios for mortality due to other causes for smokers*

Notes: all ratios defined relative with respect to smallest value within two-dimensional range.

We could have filled the matrix U element-wise instead of row-wise to find the best fitting model. However, the results have been presented to show the potential of the quadratic hazard model rather than definite results. The U-shaped relations found agree with literature, especially for BMI.

# 5. Discussion and conclusions

### The change of risk factors over time and age

At first we have used some simple methods to analyse the risk factor data from the Zutphen-Study to look for trends over age and time (see §3.2). The scatter plots for time and age do not suggest any clear trend of systolic bloodpressure, cholesterol level and Body Mass Index. The Box plots and correlation coefficients suggest changes over time and/or age for several risk factors. However, the Box-plots and correlations have to be interpreted carefully due to non-ignorable missing values (right censoring). In other words, the change of the population risk factor probability distribution over time and age is not the same as the one for any individual. The model of Manton&Stallard enabled us to analyse the change of population risk factor levels in relation to mortality and deterministic and random changes within individuals.

### Survival and cause-specific mortality

The causes of death that have been distinguished are only broad categories. The selection with respect to vascular diseases was meaningful in view of the selected set of risk factors. The distinction between lung cancer and other cancers was maybe less meaningful. The Kaplan-Meier estimations of the survival functions showed a more typical ‘survival function’ form when conditioned on age than when conditioned on time. This was caused by population heterogeneity with respect to age. In 1985 almost 50% of the population has died, in 1990 almost 65%. After some initial fluctuations the mortality proportions are almost constant over age. The proportions for CHD and cancer decrease slightly, the proportion for the other causes of death increases slightly. The proportions being almost constant over age supports the proportional hazards assumption of the mortality models used.

### The hazard function

Both the hazard function of the Cox model and the one used in the Manton&Stallard model are made functionally dependent on a set of explanatory variables through a regression function. In the Cox model a loglinear regression function is chosen, in the Manton&Stallard model a quadratic function. The different forms of the regression functions make it hard to compare the parameters. A linear instead of quadratic hazard ratio increases the comparability, but has one major objection: positiveness of the hazard function is not guaranteed any more.

### The model results with respect to the risk factor changes

Several parameterisations of the model part on the change of the risk factor levels over time have been analysed. These parameterisations differ in the interactions between the explanatory variables included. The interactions between the deterministic change have been described by the matrix $A_1$, those between the random changes by matrix D. The matrices $A_1$ and D have been assumed full (admitting all interactions) or diagonal (admitting only linear deterministic and constant random changes).

The most striking result was that the model showed increasing risk factor levels for all risk factors (although non-significant for cholesterol), whereas a simple regression analysis resulted in decreasing levels for cholesterol and BMI. Apparently too simple models are unable to take account of the complex processes underlying age-trends. We have taken account of the aspects of deterministic changes (drift, trend) versus random changes (diffusion), and interactions between the deterministic and random changes. We have not analysed the effect of time ('period-effects') apart from age so far.

For all risk factors the deterministic changes were negatively associated with the absolute values, i.e. high levels tend to increase less than low levels. This interaction was smallest for BMI. For each risk factor the deterministic changes were most strongly related with the absolute values for that risk factor and less with those for the other risk factors. This result especially applied to BMI. For bloodpressure and cholesterol there was a small interaction with BMI. The random changes found were relatively large compared to the 1-year deterministic changes, especially for SBP and cholesterol. The relative random changes were smallest for BMI, as we already had expected. The interpretation of the random changes was not clear (see below). Starting from the baseline model describing only interactions 'within' each risk factor, including interactions between the random changes (i.e. extending the matrix D to a full matrix) improved the model fit more than including interactions between the deterministic changes (i.e. extending the matrix $A_1$). Although the nature of the random changes is not clear (diffusion or measurement errors), the latter result showed that these random changes are important and are correlated.

When fitting the model only on data until 1970 the assumption of time-independent parameters seems to be valid, and probably no significant time trends could have been found. Fitting the model on all data (until 1985) would provide information on time trends. However, the large differences of the measurement time intervals (1 year until 1970 and since then 7 and 8 years respectively) together with the relatively large number of missing values after 1970 questions the usefulness of fitting the model on the entire data set. The age trends that we have found were of the same order as those found in the Monitoring Project (Verschuren et al., 1994). The differences are probably caused by the different time periods, 1960-1970 versus 1987-1991.

### The model results with respect to mortality

The cause-specific baseline hazard functions have been assumed proportional in all analyses. For the Cox model we have found that non-proportional cause-specific baseline hazard functions did not result in significant better model fits compared to proportional ones. The assumption of an exponential autonomous change over age instead of a non-parametric one (as in the Cox model) seems to be stricter, although it is generally agreed that for ages not too low or too high the exponential model is valid,

The main significant parameters are related to systolic bloodpressure (with respect to heart diseases mortality, mortality due to other causes and total mortality), to serum cholesterol (CHD mortality), Body Mass Index (lung cancer mortality and mortality due to other causes), and smoking (for total and lung cancer mortality). We have found that using current instead of baseline risk factor values changed the significance of the relations. For example, bloodpressure becomes non-significant for CHD and CVA mortality, smoking for total and CHD mortality. However, BMI becomes significant for lung cancer mortality. In the latter case the mechanism of causation is probably reverse.

The hazard function used in the Manton&Stallard model contains a quadratic regression function that describes the functional relationship to the explanatory variables. It was impossible to compare the model fits of the Cox model and the Manton&Stallard model (in terms of loglikelihood values). The reason was that in the former case the autonomous change over age was included in the model, while in the latter case the data were transformed to adjust for this change. The results on mortality agreed with those from Cox mortality analyses, qualitatively and quantitatively. The results showed significant U-shaped risk curves for cholesterol for CHD mortality and non-significant U-shaped risk curves for BMI for mortality due to other causes. The formulas suggested that it is hard to compare the results of mortality analyses on different data sets, because the regression parameters highly depend on the selection of covariates. Comparison with the results that have been presented in (Manton&Stallard, 1988) supported this conjecture.

**Measurement errors**

The model results could be biased by measurement errors and biological variability. This phenomenon is called ‘regression dilution’. It describes the dilution/attenuation in a regression coefficient that occurs when a single measured value of a covariate is used instead of the mean of two or more measured values (Knuiman et al., 1998). There are two adjustment methods, one adjusting the covariates and then fitting the model on the adjusted data, the other fitting the model on the unadjusted data and then adjusting the estimated regression parameters. In our data only single measured values were available, so we formally had to take account of regression dilution. The residuals with respect to the risk factor levels found when fitting the risk factor change submodel could have been used to estimate the variation due to measurement errors and biological variations (cf. Blomqvist, 1977). However, there are also other ways of estimating this variation, e.g. external sources or combining all repeated measurements. We have not worked out the analyses taking account of the 'regression dilution' so far.

## Model applications and further research

The model can only be fit on data from longitudinal studies. Although equally spaced measurement time points are not necessary, they simplify the model equations. Because these types of data are not generally available, model applications are limited in general.

The part of the model of Manton&Stallard describing the risk factor changes distinguishing deterministic and random changes is an example of a model on repeated measurements. Zeger&Liang (1992) have described three types of models that are used to analyse longitudinal data. The model of Manton&Stallard belongs to the group of transition (Markov) models. The part of the model that describes mortality is interesting because it introduces quadratic mortality risk function in a simple way. Interest in quadratic functions to describe mortality risks is increasing, e.g. with respect to the risk factors BMI and bloodpressure. The surplus value of the model of Manton&Stallard is combining the change of risk factor levels and mortality in one model structure. For example, Manton et al (1993) have used this model to estimate the life expectancy that could be achieved by slowing down the increase of the risk factor levels. This life expectancy can be interpreted as a 'maximum life expectancy'.

At first sight the Manton&Stallard model seemed strongly similar to other types of demographic-epidemiological simulation models in public health such as Prevent (Gunning-Schepers, 1988), TAM (Barendregt&Bonneux, 1998), CZM (Hoogenveen et al., 1998) and POHEM (Wolfson, 1991). However, we found also some major differences. The Manton&Stallard model and model results can be used to further develop these types of models. For example, the modelling of the change of specific risk factor levels could be improved by including interactions between these changes.

# References

Barendregt JJM, Bonneux L. Degenerative disease in an aging population. Models and conjectures. Thesis. IMGZ, Erasmus University Rotterdam, 1998

Becker RA, Chambers JM, Wilks AR. The new S Language. A programming environment for data analysis and graphics. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California, 1988

Blomqvist N, On the relation between change and initial value. Journal of the American Statistical Association 72 (1977), pp 746-749

Cupples LA, dÁgostinho RB, Anderson K, Kannel WB. Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study. Statistics in Medicine 7 (1988), pp 205-218, with discussion

Feskens EJM. Epidemiological studies on glucose tolerance in relation to dietary determinants and cardiovascular risks. Thesis. Leiden University, 1991

Gunning-Schepers LJ. The health benefits of prevention, a simulation approach. Thesis. IMGZ, Erasmus University Rotterdam, 1988

Hoogenveen RT, Feskens EJM, Heisterkamp SH, Lezenne Coulander C de, Bloemberg BPM. Results of an analysis of the Zutphen-Study with respect to competing death risks. [In Dutch] Report nr. 44111101. RIVM, Bilthoven, 1993

Hoogenveen RT, Jager JC. Survival analysis and competing death risks. An introduction. [In Dutch] Report nr. 958606 001. RIVM, Bilthoven, 1990

Hoogenveen RT, Hollander AEM de, Genugten MLL van. The chronic diseases modelling approach. Report nr. 266750 001. RIVM, Bilthoven, 1998

Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. John Wiley & Sons, New York [etc.], 1980

Keys A. Seven countries: a multivariate analysis of death and coronary heart disease. Harvard University Press, Cambridge, 1980

Knuiman MW, Divitini ML, Buzas JS, Fitzgerald PEB. Adjustment for regression dilution in epidemiological regression analyses. Annals of Epidemiology 8: 1 (1998), pp 56-63

Kromhout D, Bosschieter EB, De Lezenne Coulander C. Dietary fibre and 10-year mortality from coronary heart disease, cancer and all causes. The Zutphen Study. The Lancet 8297 (1982), pp 518-522

Manton KG, Stallard E. Chronic disease modelling: Measurement and evaluation of the risks of chronic disease processes. Charles Griffin & Company Ltd., Oxford; Oxford University Press, New York [etc.], 1988

Manton KG, Singer BH, Suzman RM. Forecasting the health of elderly populations. Springer Verlag, New York [etc.], 1993

Manton KG, Woodbury MA, Stallard E. Models of the interaction of mortality and the evolution of risk factor distribution: a general stochastic process formulation. Statistics in Medicine 7 (1988), pp 239-256, with discussion

Menotti A, Kromhout D, Nissinen A, Giampaoli S, Seccareccia F, Feskens E, Pekkanen J, Tervehauta M. Short-term all cause mortality and its determinants in elderly male populations in Finland, the Netherlands, and Italy; the Fine Study. Preventive Medicine 25 (1996) 3

Mulder PGH. The simultaneous processes of ageing and mortality. Statistica Neerlandica 47 (1993), pp 253-67

Uffink GJM. Analysis of dispersion by the random walk method. Thesis. University Delft, Delft, 1990

Verschuren WMM, Smit HA, Leer EM van, Berns MPH, Blokstra A, Steenbrink-van Woerden JA, Seidell JC. Prevalence of cardiovascular risk factors and changes in risk factors over the period 1987-1991. Final report on the Monitoring Project on Cardiovascular Disease Risk Factors. [In Dutch] Report nr. 528901 011. RIVM, Bilthoven, 1994

Voedingsraad. Epidemiologic research ‘Nutrition and atherosclerosis in Zutphen’. [In Dutch] Staatsuitgeverij, ‘s Gravenhage, 1984

Wolfson MC. A system of health statistics toward a new conceptual framework for integrating health data. Review Of Income and Wealth 37: 1 (1991), pp 81-104

Yashin AI, Manton KG, Stallard E. Dependent competing risks: a stochastic process model. Journal of Mathematical Biology (1986), vol 24, pp 119-40

Zeger SL, Liang K-Y. An overview of methods for the analysis of longitudinal data. Statistics in Medicine 11 (1992), pp 1825-1839

# Appendix 1 Mailing list

| | |
|---|---|
| 1 | Directeur-Generaal RIVM |
| 2 | Dr HJ Schneider, Directeur-Generaal van de Volksgezondheid |
| 3 | Prof dr JJ Sixma, Voorzitter van de Gezondheidsraad |
| 4 | Drs PH Vree, waarnemend Hoofdinspecteur voor de Gezondheidszorg |
| 5 | Dr JJM Barendregt (EUR) |
| 6 | Dr SH Heisterkamp (AMC-UvA) |
| 7 | Dr BA van Hout (EUR) |
| 8 | Dr PGH Mulder (EUR) |
| 9 | Drs LW Niessen (EUR) |
| 10 | Depot Nederlandse Publikaties en Nederlandse Bibliografie |
| 11 | Dr HC Boshuizen |
| 12 | Dr EJM Feskens |
| 13 | Ir MGG van Genugten |
| 14 | Ir P van den Hoogen |
| 15 | Drs S Houterman |
| 16 | Ir J Jansen |
| 17 | Prof dr ir D Kromhout |
| 18 | Ir I Mulder |
| 19 | Prof dr ir JC Seidell |
| 20 | Dr ir WMM Verschuren |
| 21 | Ir TLS Visscher |
| 22 | Auteur |
| 23 | SBD/Voorlichting & Public Relations |
| 24 | Bureau Rapportenregistratie |
| 25 | Bibliotheek RIVM |
| 26-37 | Bureau Rapportenbeheer |
| 38-50 | Reserve exemplaren |

# Appendix 2 Formal proofs of some results

We have provided some formal proofs of the main steps of the development of the Manton&Stallard model. The derivation of the Kolmogorov-Fokker-Planck differential equation can be found in literature, e.g. Uffink (1990).

**The 'mortality part' of the Kolmogorov-Fokker-Planck partial differential equation**

$f(z,\tau>t+\Delta t|\tau>t) = f(z|\tau>t+\Delta t) * \Pr(\tau>t+\Delta t|\tau>t) = \Pr(\tau>t+\Delta t|\tau>t,z) * f(z|\tau>t)$

with: $f(z,\tau>t+\Delta t|\tau>t)$: the joint probability density function of the explanatory variables and time of death.

$=>$ $\Delta f(z|\tau>t) = f(z|\tau>t+\Delta t) - f(z|\tau>t) = f(z|\tau>t) \{ \Pr(\tau>t+\Delta t|\tau>t,z)/\Pr(\tau>t+\Delta t|\tau>t) - 1 \}$

$= f(z|\tau>t) \{ (1-\mu(t,z)\Delta t)/(1-\mu(t)\Delta t) - 1 \}$

$= f(z|\tau>t) \{ (1-\mu(t,z)\Delta t) - (1-\mu(t)\Delta t) \} / \{ 1-\mu(t)\Delta t \}$

$=>$ $d/dt\, f(z|\tau>t) = f(z|\tau>t) \{ \mu(t) - \mu(t,z) \}$

**The population hazard function**

$\mu(t) = E( \mu(t,z_t) \mid \tau>t ) = E( b_0 + b_1'z_t + ½z_t'Bz_t \mid \tau>t )$

$= b_0 + b_1'E(z_t|\tau>t) + ½\{ E(z_t|\tau>t)' B E(z_t|\tau>t) + tr( V_tB ) \}$

$= \mu(t,m_t) + ½ tr( V_tB )$

**The ordinary differential equations for the mean and variance of the multivariate normal probability density function**

(1) Formal differentiation of the density function of the multivariate normal distribution:

$f_t(z) = (2\pi)^{-½n} |V_t|^{-½} \exp\{ -½(z-\mu_t)'V_t^{-1}(z-\mu_t) \}$

The time-derivative of a matrix determinant is:

$V_t^{-1} V_t = I,$ $\quad adj(V_t) = |V_t| V_t^{-1}$

$$\Rightarrow \quad \mathrm{tr}(\ V_t^{-1}\ dV_t/dt\ ) = |V_t|^{-1}\ \mathrm{tr}(\ \mathrm{adj}(V_t)\ dV_t/dt\ ) = |V_t|^{-1}\ d|V_t|/dt$$

Thus: $\delta/\delta t \ln f_t(z) = -½n(\ln 2\pi) - ½\ d|V_t|/dt\ /\ |V_t| - ½(z-\mu_t)'V_t^{-1}(z-\mu_t)$

(2) Substituting the deterministic and random changes in the Kolmogorov-Fokker-Planck partial differential equation gives:

$$\delta/\delta t\ f_t(z) = -\ \Sigma_j\ \{\ \delta/\delta z_j\ [\ u_j(z)*f_t(z)\ ]\ \} + ½\ \Sigma_{i,j}\ \{\ \delta^2/\delta z_i\delta z_j\ [\ \sigma_{ij}\ f_t(z)\ ]\ \} - \{\ \mu(t,z)-\mu(t)\ \}\ f_t(z)$$

$$\delta/\delta t \ln f_t(z) = \delta/\delta t\ f_t(z)\ /\ f_t(z)$$

$$= -\ \Sigma_j\ \{\ u_j(z)\ \delta/\delta z_j \ln f_t(z)\ \} - \Sigma_j\ \{\ \delta/\delta z_j\ u_j(z)\ \}\ +$$

$$½\ \Sigma_{i,j}\ \{\ \sigma_{ij}\ [\ \delta^2/\delta z_i\delta z_j \ln f_t(z) + \delta/\delta z_i \ln f_t(z) * \delta/\delta z_j \ln f_t(z)\ ]\ \} - \{\ \mu(t,z)-\mu(t)\ \}$$

with:

$u_j(z) = (\ A_0 + A_1 z\ )_j \quad \Rightarrow \quad \delta/\delta z_j\ u_j(z) = [A_1]_{jj}$

$\delta/\delta z_j \ln f_t(z) = V_t^{-1}(z-\mu_t)_j \quad \Rightarrow \quad \delta^2/\delta z_i\delta z_j \ln f_t(z) = [V_t^{-1}]_{ij}$

$\mu(t,z) = b_0 + b_1'z + ½z'Bz$

$\Rightarrow \delta/\delta t \ln f_t(z) =$

$$(A_0+A_1z)'V_t^{-1}(z-m_t) - \mathrm{tr}(A_1) - ½\mathrm{tr}(\Sigma V_t^{-1}) + ½(z-m_t)'V_t^{-1}\Sigma V_t^{-1}(z-m_t)\ +$$

$$-\ b_0-b_1'z-½z'Bz + \mu(t)$$

The resulting formulas from (1) and (2) should be identical functions of the variable z. So we equate the terms that are quadratic in z, resulting in the differential equation for $V_t$, we equate the terms that are linear in z, resulting in the differential equation for $m_t$, and we equate the terms that are independent on z, resulting in the equation for $\mu(t)$.

**The discrete approximation step related to mortality**

We use the same equation that underlies the mortality (survival) selection part of the Kolmogorov-Fokker-Planck partial differential equation:

$$f(z|\tau>t+\Delta t) * \Pr(\tau>t+\Delta t|\tau>t) = \Pr(\tau>t+\Delta t|\tau>t,z) * f(z|\tau>t)$$

Substituting the multivariate normal variable probability density function yields:

$$(2\pi)^{-n/2}\,|V_{t+1-}|^{-½}\,\exp\{\,-½(z-m_{t+1-})'V_{t+1-}^{-1}(z-m_{t+1-})\,\}*\exp(\,-\mu(t)\,)=$$

$$(2\pi)^{-n/2}\,|V_t|^{-½}\,\exp\{\,-½(z-m_t)'V_t^{-1}(z-m_t)\,\}*\exp\{\,-(b_0+b_1'z+z'Bz)\,\}$$

The right and left hand side of the equation are identical functions of the variable z. So we again equate the terms that are quadratic in z, resulting in the equation for $V_{t+1-}$, we equate the terms that are linear in z, resulting in the equation for $m_{t+1-}$, and we equate the terms that are independent on z, resulting in the equation for $\mu(t)$. Then it follows:

$$V_{t+1-}^{-1}=V_t^{-1}+B=V_t^{-1}\,(I+V_tB)$$

$$\Rightarrow\qquad V_{t+1-}=D_tV_t$$

$$V_{t+1-}^{-1}m_{t+1-}=V_t^{-1}m_t-b_1$$

$$\Rightarrow\qquad m_{t+1-}=V_{t+1-}(V_t^{-1}m_t-b_1)=D_tV_tV_t^{-1}(m_t-V_tb_1)=D_t\,(m_t-V_tb_1)$$

with: $D_t=(I+V_tB)^{-1}$. According to (Manton&Stallard, 1988):

$$m_{t+1-}=m_t-D_tV_t\,(b_1+Bm_t)=(I-D_tV_tB)\,m_t-D_tV_tb_1$$

$$I-D_tV_tB=D_t\,(D_t^{-1}-V_tB)=D_t\,(I+V_tB-V_tB)=D_t$$

**The discrete approximation step related to ageing**

$$z_{t+1}=z_{t+1-}+dz_{t+1-}=z_{t+1-}+(\,A_0+A_1z_{t+1-}\,)dt+D^Tdw_t$$

Thus:

$$m_{t+1}=E(z_{t+1})=Ez_{t+1-}+E\{\,(A_0+A_1z_{t+1-})dt+D^T\,dw_t\,\}$$

$$=m_{t+1-}+A_0+A_1\,Ez_{t+1-}+D^T\,0=A_0+(\,I+A_1\,)\,m_{t+1-}$$

$$V_{t+1}=\mathrm{var}(z_{t+1})=\mathrm{var}(\,A_0+(I+A_1)z_{t+1-}\,)+\mathrm{var}(\,D^T\,dw_t\,)$$

$$=(I+A_1z_{t+1-})\,\mathrm{var}(z_{t+1-})\,(I+A_1z_{t+1-})+D^T\,\mathrm{var}(dw_t)\,D$$

$$=(I+A_1z_{t+1-})\,V_{t+1-}\,(I+A_1z_{t+1-})+D^T\,D$$

assuming that the deterministic and random changes are independent processes.

### The likelihood function

We have assumed that all individuals are independent. So the likelihood function is the product of similar separate likelihood functions for all individuals.

$$L( \{X_i(t_0..t_N)\},\{\tau_i\}_{i \varepsilon I} ; A,\Sigma,Q ) \propto \Pi_{i \varepsilon I} L_i( X_i(t_0..t_N),\tau_i ; A,\Sigma,Q )$$

$$L_i \propto P( X_i(t_0,t_N),\ \tau_i<\tau\leq\tau_i+\Delta t )$$

$$\approx f_0(x_{i0})\ \ f_1(x_{i1}|X_i(t_0,t_0),\tau>t_0)\ \ f_2(X_{i2}|X_i(t_0,t_1),\tau>t_1) \ldots$$

$$P(\tau>\tau_i|X_i(t_0,\tau_i))\ \ Pr(\tau\leq\tau_i+\Delta t|X_i(t_0,\tau_i),\tau>\tau_i)$$

$$\approx f_0(x_{i0})\ \ f_1(x_{i1}|x_{i0},\tau>t_0)\ \ f_2(x_{i2}|x_{i1},\tau>t_1) \ldots$$

$$\mu( [\tau_i],x_i([\tau i])\ \ P(\tau>t_0|x_{i0})\ \ Pr(\tau>t_1|,x_{i1},\tau>t_1)\ \ ..\ \ P(\tau>\tau_i|x_{i,[\tau i]},\tau>[\tau_i]))$$

$$= f_0(x_{i0})\ \ \Pi_{0<n\leq[\tau i]} f_n(x_i(t_n)|x_{i,n-1})$$

$$\mu( [\tau_i],x_i([\tau i]) )\ \ \Pi_{0\leq n<[\tau i]} \exp\{ -\mu( t_n,x_i(t_n) ) \}\ \ \exp\{ -\Theta_i\, \mu( [\tau_i],x_i([\tau i] ) \}$$

$$= f_0(x_{i0})\ \ \Pi_{0<n\leq[\tau i]} f_n(x_i(t_n)|x_{i,n-1})$$

$$\mu( [\tau_i],x_i([\tau i]) )\ \ \Pi_{0\leq n\leq[\tau i]} \exp\{ -w_{in}\, \mu(t_n,x_i(t_n)) \}$$

with: $w_{in}$: the length of the n-th subject-interval for individual i. We have used the following assumptions: (1) the Markov-property, meaning that conditional on the last variable values the past values do not affect the future mortality and variable change process; (2) the time of death is observed. In case of non-informative censoring the last term, that contains the hazard function value for the last time period, has to be omitted.

### The quadratic hazard function model in case of different death risks

We assume (for every $k \varepsilon K$) that $\mu_k(t,z_t) = z_t^{*\prime}\, Q_k\, z_t^{*}$, and that $Q_k$ is a symmetric non-negative definite matrix of bounded time-dependent coefficients. Then:

$$\mu(t,z_t)\ \Delta t \approx Pr( t<\tau\leq t+\Delta t \mid \tau>t,z_t ) \approx \Sigma_{k \varepsilon K} Pr( t<\tau\leq t+\Delta t,C=k \mid \tau>t,z_t )\ \Delta t$$

$$\approx \Sigma_{k\varepsilon K} ( z_t^{*'} Q_k z_t^{*} ) \Delta t = z_t^{*'} ( \Sigma_{k\varepsilon K} Q_k ) z_t^{*} \Delta t$$

Therefore $Q = \Sigma_{k\varepsilon K} Q_k$ and Q is still a symmetric non-negative definite matrix of bounded time-dependent coefficients.

**The likelihood function part related to mortality distinguishing different causes of death**

$$L_M \propto \Pi_{i\varepsilon I} \{ \Pi_{0\leq n\leq[\tau i]} \exp\{ -w_{in} \mu(t_n,x_i(t_n) ) \} \} \; \Pi_{k\varepsilon K} \Pi_{i\varepsilon Mk} \{ \mu_k( [\tau_i],x_i([\tau i]) ) \}$$

$$= \Pi_{i\varepsilon I} \{ \Pi_{0\leq n\leq[\tau i]} \exp\{ -w_{in} \Sigma_{k\varepsilon K} \mu_k(t_n,x_i(t_n) ) \} \} \; \Pi_{k\varepsilon K} \Pi_{i\varepsilon Mk} \{ \mu_k( [\tau_i],x_i([\tau i]) ) \}$$

$$= \Pi_{i\varepsilon I} \{ \Pi_{k\varepsilon K} \Pi_{0\leq n\leq[\tau i]} \exp\{ -w_{in} \mu_k(t_n,x_i(t_n) ) \} \} \; \Pi_{k\varepsilon K} \Pi_{i\varepsilon Mk} \{ \mu_k( [\tau_i],x_i([\tau i]) ) \}$$

$$= \Pi_{k\varepsilon K} \{ \Pi_{i\varepsilon I} \Pi_{0\leq n\leq[\tau i]} \exp\{ -w_{in} \mu_k(t_n,x_i(t_n) ) \} \; \Pi_{i\varepsilon Mk} \{ \mu_k( [\tau_i],x_i([\tau i]) ) \} \}$$

$$= \Pi_{k\varepsilon K} L_{Mk}$$

Therefore the mortality part of the likelihood function can be multiplicatively separated into independent, functionally similar cause-specific likelihood functions.

**The likelihood function part related to mortality using subject-intervals**

The likelihood function mortality part can be reformulated using the concept of subject-intervals. We assume that all times of death are observed and we distinguish no separate death risks here. The results can easily be generalised to the cases of right-censoring and distinguishing causes of death.

$$L_m \propto \Pi_{i\varepsilon I} \{ \Pi_{0\leq n\leq[\tau i]} \exp\{ -w_{in}{*}\mu(t_n,x_i(t_n) ) \} \; \mu( [\tau_i],x_i([\tau i]) ) \} = \Pi_{i\varepsilon I*} \exp\{ - w_i \mu_i \} \; \Pi_{i\varepsilon M*} \mu_i$$