Web Material to:
From atherosclerosis to myocardial infarction – a process-oriented model investigating the role of risk factors

Dr. Cristoforo Simonetto, Dr. Margit Heier, Prof. Dr. Annette Peters,
Dr. Jan Christian Kaiser, Dr. Susanne Rospleszcz

Web Appendices

Web Tables

Web Figures

Web Appendix 1 A descriptive model

A traditional descriptive model was developed for preliminary analysis and for comparison to the process-oriented model. The hazard h of this model can be written as:

$$\begin{aligned}\log h = &\beta_0 + \beta_1 \log(a) + \beta_2 \log^2(a) \\ &+ \beta_3 \Theta_{\text{active smoker}} \\ &+ \beta_4(sys - 130) + \beta_5 \Theta_{\text{antihypertensive therapy}} \\ &+ \beta_6(hdl - 1.5) + \beta_7(nonhdl - 4.5) \\ &+ \beta_8(1940 - by)\Theta_{by<1940} \end{aligned} \tag{1}$$

Here the β_i denote model coefficients, a denotes age [years], by birth year, sys systolic blood pressure [mmHg] and hdl and $nonhdl$ the HDL and non-HDL cholesterol levels [mmol/l]. The indicator function Θ equals 1 if the subscript is true and 0 otherwise.

The model was applied separately for both sexes. It was developed starting from a model that depends only on age, successively adding other risk factors. Including smoking information with an indicator function yielded a better fit compared to a model linear in cumulative pack years or a model linear in current smoking intensity. Diastolic blood pressure was not a significant predictor after inclusion of systolic blood pressure. Also antihypertensive therapy was included but not lipid medication as the latter was not statistically significant. The probable reason for the difference in significance is that 17% of the cohort received treatment for blood pressure but only 4% for lipids. Finally, linear trends in calendar year and in birth year before 1940 were tested, following [1]. However, only the trend with birth year was significant and only in men. To correct for birth year also in the process-oriented model, the same birth-year dependent factor, eq. (1), was applied, and with the same value $\beta_8 = 0.04$ that was obtained from the descriptive model.

Web Appendix 2 Imputation of the age trajectory of risk factors

Levels of risk factors were determined at examination. However, the process-oriented model utilizes annual risk factor levels. Therefore, the age dependence of risk factors was imputed as follows. Smoking was assumed to affect biological parameters only during years of active smoking. Age of smoking initiation and cessation was recorded in the S3 but not in the S4 survey. Instead, only smoking status was assessed in S4 (active, never, or former smoker). For smokers without specific information, smoking was assumed to start at age 18, the median start age in S3. Other missing information was imputed by sampling, as multiple model simulations were performed for each individual, anyway. For former smokers, age of smoking cessation was sampled uniformly between age 18 and the age of examination. For active smokers from both surveys, smoking cessation after examination was estimated based on the average cessation rates of smokers in S3: 1.3% per year for men and 1.0% for women. Persons who have smoked not more than 1 pack-year before examination were classified as non-smokers.

Linear fits of systolic blood pressure were performed to obtain sex specific cohort means and standard deviations in dependence of age at examination, see Table 1. These fits were used to convert absolute individual blood pressure readings to deviations from the mean, given in units of standard deviations. For example, a woman at age 65 may present with a systolic blood pressure of 150 mmHg. The mean pressure in the cohort is 138 mmHg at age 65, and the standard deviation 21 mmHg. Therefore, her blood pressure is $(150 - 138)/21 = 0.57$ standard deviations above the mean. In the absence of antihypertensive therapy, these deviations were assumed to be constant throughout life. At age 55, she is therefore attributed 141 mmHg. Antihypertensive therapy was assumed to permanently reduce systolic blood pressure (deviation) by some value to be determined by the fit, but at most two standard deviations. Information on medication was recorded in the survey, but not the age of start of antihypertensive therapy. Therefore, start of medication was imputed based on the prevalence of antihypertensive therapy in the cohort. In persons examined before age 40, only 38 persons were treated for hypertension. Thereafter, the proportion increases and reaches 40% for men above age 65 and 47% for women. Therefore, first treatment age was sampled uniformly between age 40 and the age of examination. For persons not treated at the age of examination, treatment was assumed to begin with a probability of 3.9% for men and 8.3% for women in each year after examination with estimated systolic blood pressure above 140 mmHg. These numbers were obtained by first calculating for each age at examination the share of persons treated for hypertension and the share of persons not treated despite a systolic blood pressure above 140 mmHg. Then, the mean share of persons treated, examined at ages above 70, was divided by the sum of the shares not treated at younger ages.

Both, HDL and non-HDL cholesterol levels were implemented in the model. Analogous to blood pressure, values were converted to deviations to the age- and sex-specific cohort mean, and these deviations were assumed to be independent of age, to derive estimates for cholesterol levels during the whole life.

Web Appendix 3 Details to the fitting procedure

Apart from their dependence on risk factor levels, most parameters in the model were assumed constant. For the growth rate of complicated lesions γ_3, and the risk of occlusion at the site of the complicated lesion leading to infarction ν_h, however, application of the process-oriented model to registry data had revealed an age dependence for women during menopause [1]. This age dependence could neither be verified nor falsified in the present study, given the lower case numbers. Therefore, the same age dependence was adopted which implies a rise in the growth rate of complicated lesions from $0.37\gamma_3$ to γ_3 during menopause and a decline of the rate of occlusion from $3\nu_h$ to ν_h. To further reduce the complexity of the model, also the size of newly initiated lesions s was fixed, $s = 0.85\%$ of the artery intimal surface area for men and $s = 0.82\%$ for women [1].

Best estimates for all other biological parameters (transition and growth rates, as well as the risk factor dependencies thereof) were obtained by minimizing the total deviance, given by the sum of the deviance dev_{epi} related to the epidemiological survey data and the deviance dev_{sub} related to the subclinical data. The deviance dev_{epi} related to the epidemiological data was calculated from the individual likelihoods l_i by $dev_{epi} = -2\log(\prod l_i)$. For each member i of the epidemiological data, a hazard $h_i(a)$ and related survival function $S_i(a) = \exp(-\int_0^a h_i(t)dt)$ was calculated for each age a. As usual, the individual likelihood,

$$l_i = h_i(a_{out})^{\delta_i} S_i(a_{out})/S_i(a_{in})$$

represents the likelihood of the observed outcome: $\delta_i = 1$ for persons with MI incidence at age a_{out} and $\delta_i = 0$ for persons with no incidence to the end of follow up at age a_{out}. Age at study entry is a_{in}.

To calculate the hazard h_i of each single cohort member, 1000 hazard functions $h_{i,z}$ $(z = 1 \dots 1000)$ were simulated by the model and weighted with the corresponding survival:

$$h_i(a) = \frac{\sum_{z=1}^{1000} h_{i,z}(a) S_{i,z}(a)}{\sum_{z=1}^{1000} S_{i,z}(a)} \tag{2}$$

The same simulations were also used to calculate the deviance dev_{sub} related to subclinical data. For this purpose, a total number of 10,000 samples was collected randomly from the simulated data. Each sample consists of 212 individuals – the number of persons contained in the subclinical data set. Moreover, samples were chosen to reflect the age and sex structure of the data. The same measures of lesion spread were calculated from each simulation sample that were also presented in the subclinical data: prevalence, prevalence of significant lesions, mean lesion area, and standard deviation of the lesion area. The likelihood of observing the subclinical data given the model is then read off from the distribution of the 10,000 results. The deviance is given by twice the negative logarithm of the likelihood. Details are described in [1].

Uncertainty due to the finite number of samples is assessed below. To examine whether the epidemiological data is well described by the process-oriented model, only dev_{epi}, the part of the deviance related to the epidemiological data, is compared to the deviance of the descriptive model in Web Appendix 5.

Web Appendix 4 Uncertainty in the deviance calculation

The deviance is used as a measure of goodness-of-fit to discriminate between model variants. However, due to the finite number of simulations of the process-oriented model, inevitably there is some uncertainty in the deviance calculation. To assess this uncertainty, the final model with cholesterol affecting ν_1 (parameters in Table 3 in the main text) and the model with cholesterol affecting ν_2 were both evaluated 10 times using different random seeds. In the main analysis, these two models differed by 5.1 points in the deviance for men, and 3.7 for women, see Table 2 in the main manuscript. Using different random seeds, the deviance differences varied with a standard deviation of 1.2 and 1.3 respectively. While such uncertainty is not negligible, it does not compromise our main results which are based on larger deviance differences and comparison to early lesion development.

Web Appendix 5 Benchmarking with a descriptive model

The descriptive model, eq. (1), was used to benchmark the process-oriented model. For each risk factor, a preferred model variant of the process-oriented model was established in the main manuscript. The part of the deviance dev_{epi} that is related to the epidemiological data, is compared to descriptive model fits in Table 2. Both, descriptive and process-oriented model equally well describe baseline risk as well as any risk factor in men. For women, inclusion of blood pressure tends to be better described by the process-oriented model. It should be noted, however, that deviance evaluation in the process-oriented model is associated with uncertainties, see the last section.

Web Appendix 6 Hazard ratios

The marginal hazard h_i describes the average risk given the known risk factors pertinent to an individual. It is calculated by simulating possible hazard functions $h_{i,z}$ and weighting them with the corresponding survival function, see eq. (2) Each simulated hazard $h_{i,z}$ presents a possible risk of an individual described by risk factors i. The weighting reflects the fact that after first incidence individuals drop out of an incidence cohort. However, as a consequence of this weighting, hazard ratios based on the marginal hazard,

$$HR_{i,j}^{\text{marginal}} = \frac{h_i}{h_j} = \frac{\sum_z h_{i,z} S_{i,z}}{\sum_z h_{j,z} S_{j,z}} \times \frac{\sum_z S_{j,z}}{\sum_z S_{i,z}} \tag{3}$$

may decrease with age even if $h_{i,z}$ and $h_{j,z}$ differ only by a constant factor. This fact is well-known in frailty analysis [2]. In frailty models, the so called conditional hazard ratio $h_{i,z}/h_{j,z}$ is therefore applied which is conditional on known and unknown risk factors [3]. Typically, the unknown risk factors are modeled by a latent factor which drops out when performing the ratio. In that case, the conditional hazard ratio depends only on the known risk factors. However, this is not the case for our simulation approach. In the model, $h_{i,z}$ results from a complex interplay of known risk factors of the individual i, and individual growth rates and random processes reflected by the simulation z. A hazard ratio $h_{i,z}/h_{j,z}$ may be calculated by simulation for two levels of risk factors i and j. But even if simulations for $h_{i,z}$ and $h_{j,z}$ are performed with the same random numbers, the hazard ratio is meaningless: For many simulations small differences in risk factors may have no consequence, in particular if no complicated lesion is formed anyway. However, the ratio approaches infinity if a complicated lesion is formed only for i but not for j. To deal with this problem, we define a mean conditional hazard ratio:

$$HR_{i,j}^{\text{mean cond}} = \frac{\sum_z h_{i,z} S_{i,z}}{\sum_z h_{j,z} S_{i,z}} \tag{4}$$

If $h_{i,z}/h_{j,z}$ was independent of z, as usual in frailty analysis, the definition would collapse to the conditional hazard ratio $h_{i,z}/h_{j,z}$. On the other hand, eq. (4) resembles the marginal hazard ratio with the only difference of same weights in nominator and denominator. The mean conditional hazard ratio describes the fold in risk for an average individual with risk factors i compared to risk factors j. It answers an individual's question on how risk would be different if he or she had other risk factors. In contrast, the marginal hazard ratio describes the ratio of risks of an average individual with risk factors i compared to an average individual with risk factors j. As averages are performed separately, the impact of individuals with high risk may be suppressed more strongly in the nominator than in the denominator. Only the marginal hazard ratio is directly accessible by incidence data while the mean conditional hazard additionally requires information on heterogeneity of risk within individuals with the same risk factors. In Fig. 1 only the marginal hazard is shown but both, marginal and mean conditional hazard are presented in Fig. 4 in the main text.

References

[1] C Simonetto, S Rospleszcz, M Heier, C Meisinger, A Peters, and J C Kaiser. Simulating the dynamics of atherosclerosis to the incidence of myocardial infarction, applied to the KORA population. *Stat Med*, 40:3299–3312, 2021.

[2] O O Aalen. Effects of frailty in survival analysis. *Stat Methods Med Res*, 3(3):227–43, 1994.

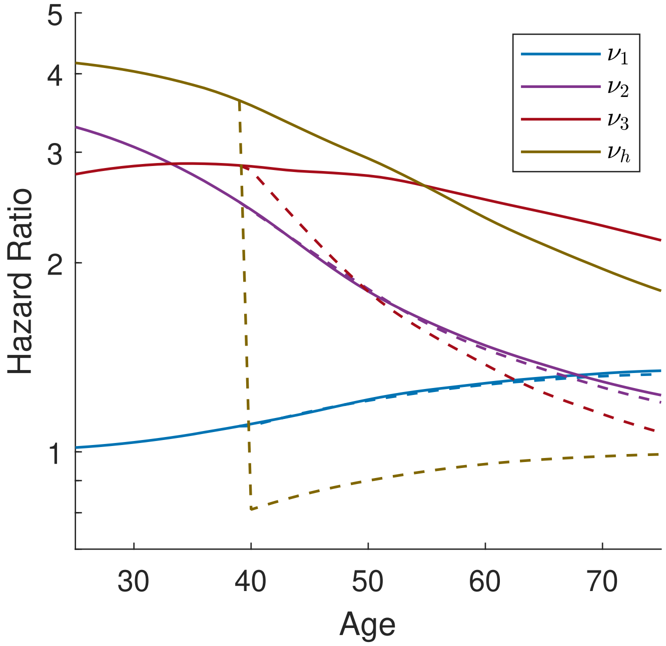
[3] T A Balan and H Putter. A tutorial on frailty models. *Stat Methods Med Res*, 29(11):3424–3454, 2020.

	Systolic blood pressure, mmHg	HDL, mmol/l	Non-HDL cholesterol, mmol/l
Men	$\mu = 112 + 0.46a$	$\mu = 1.2 + 0.001a$	$\mu = 3.9 + 0.015a$
	$\sigma = 5.6 + 0.22a$	$\sigma = 0.31 + 0.001a$	$\sigma = 1.2 - 0.001a$
Women	$\mu = 89 + 0.76a$	$\mu = 1.7 - 0.003a$	$\mu = 2.5 + 0.036a$
	$\sigma = 4.0 + 0.26a$	$\sigma = 0.42 + 0.000a$	$\sigma = 0.84 + 0.004a$

Web Table 1: Means (μ) and standard deviations (σ) of systolic blood pressure, HDL and non-HDL cholesterol as obtained from a linear fit on age (a) to the KORA survey data.

	Men		Women	
	Descriptive model	Process-oriented model	Descriptive model	Process-oriented model
baseline	3728.4	3728.5	1803.3	1805.1
+ smoking	3686.6	3688.3	1774.9	1775.2
+ blood pressure	3661.9	3660.1	1740.2	1733.1
+ lipid levels	3641.7	3641.4	1729.3	1724.1

Web Table 2: Deviance related to the epidemiological KORA survey for the descriptive and the preferred process-oriented models. In each line an additional risk factor is additionally taken into account into the models.



Web Figure 1: Hazard ratios of men who started smoking at age 18 versus non-smokers according to the process-oriented model. Each colored line corresponds to a model variant with one specific parameter depending on smoking. Dashed lines illustrate the modeled course for smoking cessation at age 40.