

# Simulating the dynamics of atherosclerosis to the incidence of myocardial infarction, applied to the KORA population

Cristoforo Simonetto<sup>1</sup>  | Susanne Rospleszcz<sup>2,3,4</sup> | Margit Heier<sup>2,5</sup> |  
Christa Meisinger<sup>6,7,8</sup> | Annette Peters<sup>2,3,4</sup> | Jan Christian Kaiser<sup>1</sup>

<sup>1</sup>Institute of Radiation Medicine, Helmholtz Zentrum München German Research Center for Environmental Health (GmbH), Munich, Germany

<sup>2</sup>Institute of Epidemiology, Helmholtz Zentrum München German Research Center for Environmental Health (GmbH), Munich, Germany

<sup>3</sup>Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany

<sup>4</sup>German Center for Cardiovascular Disease (DZHK), Partner Site Munich Heart Alliance, Munich, Germany

<sup>5</sup>KORA Study Centre, University Hospital of Augsburg, Augsburg, Germany

<sup>6</sup>MONICA/KORA Myocardial Infarction Registry, University Hospital of Augsburg, Augsburg, Germany

<sup>7</sup>Chair of Epidemiology, Ludwig-Maximilians-Universität München, UNIKA-T, Munich, Germany

<sup>8</sup>Independent Research Group Clinical Epidemiology, Helmholtz Zentrum München German Research Center for Environmental Health (GmbH), Munich, Germany

## Correspondence

Cristoforo Simonetto, Institute of Radiation Medicine, Helmholtz Zentrum München German Research Center for Environmental Health (GmbH),

Analyzing epidemiological data with simplified mathematical models of disease development provides a link between the time-course of incidence and the underlying biological processes. Here we point out that considerable modeling flexibility is gained if the model is solved by simulation only. To this aim, a model of atherosclerosis is proposed: a Markov Chain with continuous state space which represents the coronary artery intimal surface area involved with atherosclerotic lesions of increasing severity. Myocardial infarction rates are assumed to be proportional to the area of most severe lesions. The model can be fitted simultaneously to infarction incidence rates observed in the KORA registry, and to the age-dependent prevalence and extent of atherosclerotic lesions in the PDAY study. Moreover, the simulation approach allows for non-linear transition rates, and to consider at the same time randomness and inter-individual heterogeneity. Interestingly, the fit revealed significant age dependence of parameters in females around the age of menopause, qualitatively reproducing the known vascular effects of female sex hormones. For males, the incidence curve flattens for higher ages. According to the model, frailty explains this flattening only partially, and saturation of the disease process plays also an important role. This study shows the feasibility of simulating subclinical and epidemiological data with the same mathematical model. The approach is very general and may be extended to investigate the effects of risk factors or interventions. Moreover, it offers an interface to integrate quantitative individual health data as assessed, for example, by imaging.

## KEYWORDS

atherosclerosis, Markov chain, mechanistic model, myocardial infarction, simulation

**Abbreviations:** KORA, Cooperative Health Research in the Augsburg Region; PDAY, Pathobiological Determinants of Atherosclerosis in Youth.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Ingolstädter Landstraße 1, 85764  
Neuherberg, Munich, Germany.

Email:

cristoforo.simonetto@helmholtz-muenchen.de

### Funding information

H2020 Euratom, Grant/Award Number:  
755523

## 1 | INTRODUCTION

Multi-state models can be applied to describe individual health by a stochastic succession of, typically a few, possible states.<sup>1,2</sup> In cancer epidemiology, they have long tradition<sup>3</sup> and constitute a mainstay of radiation epidemiology.<sup>4</sup> In modeling carcinogenesis, cells are assumed to evolve independently of each other. This simplification allows for analytical solution of the involved stochastic differential equations and facilitates model fitting to epidemiological data.<sup>5,6</sup> Also for the distribution of pre-cancerous cells, analytical expressions have been derived<sup>7</sup> but there is still no experimental data for direct verification of the predicted evolution of cells.

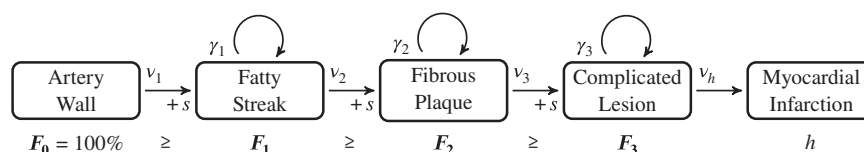
Despite the virtues of an analytical solution, here we aim to demonstrate that analytical solvability is not a necessary requirement, and abandoning it allows for versatile modeling possibilities which can only be started to be explored in the present study. To make the case, a model of atherosclerosis and subsequent myocardial infarction is presented. Compared to cancer, we believe there is an even greater potential for process oriented disease and risk modeling for cardiovascular diseases. One important reason is the possibility of directly observing the disease development at various stages. Moreover, process oriented modeling may help to better evaluate the potentially individually different short- and long-term benefits of the manifold options of prevention. Still, little effort has been put so far into the modeling of atherosclerosis development and its relation to incidence data,<sup>8,9</sup> which may partially be caused by model limitations.

Our model is solved by computer simulation. This allows to include more complex interactions, to apply a continuous state space, and to easily take into account inter-individual heterogeneity as an important source of variability of individual cardiovascular risk. Higher modeling flexibility can only be exploited with appropriate data on the modeled processes. Here, the simulation approach is advantageous by easing the inclusion of additional, subclinical data in model development and fitting. This ensures a quantitatively realistic model behavior for early, subclinical disease development even though the complex disease interactions cannot be fully represented by any model.

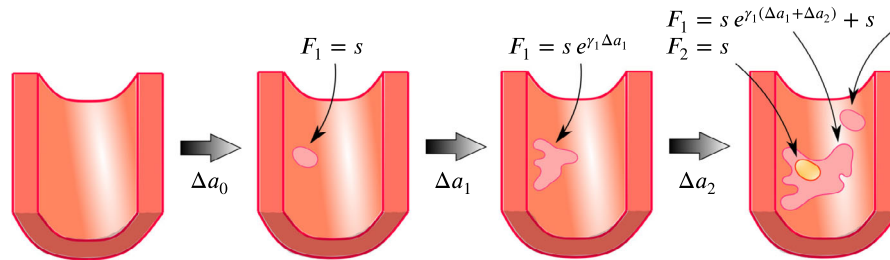
## 2 | MATERIALS AND METHODS

### 2.1 | The model

Atherosclerosis is characterized by the progressive accumulation of lipids and fibrous elements in the walls of large arteries. Lipid deposition starts at preferred sites and results in atherosclerotic lesions.<sup>10</sup> With time, the area involved with lesions increases and lesion appearance progresses. Therefore, sophisticated classification schemes of atherosclerotic lesions were developed.<sup>11</sup> Guided by the data used for the fit (see below),<sup>12</sup> lesions are classified into only three categories in the present study: fatty streaks as very early lesions, fibrous plaques, and, finally, complicated lesions.



**FIGURE 1** Schematic model representation. Percentages of the coronary artery intimal surface area involved with lesions of at least type  $k$  are denoted by  $F_k$ . This means  $k=1$  refers to all lesions,  $k=2$  to fibrous plaques or complicated lesions, and  $k=3$  to complicated lesions only. Existing lesions grow with rate  $\gamma_k$ . New lesions of type  $k$  emerge from  $F_{k-1}$  according to a Poisson process with rate  $v_k$  and with size  $s$ . Analogously, complicated lesions give rise to myocardial infarction with rate  $v_h$



**FIGURE 2** Sketch of possible early plaque development in the model. A first fatty streak appears after time  $\Delta a_0$  with area  $F_1 = s$ . Subsequently it expands with growth rate  $\gamma_1$ . At the last depicted time point, part of the fatty streak has become a fibrous plaque, with area  $F_2 = s$ . The intimal surface area involved with fatty streaks (and more advanced lesions),  $F_1$ , results from further growth of the first fatty streak and from the origin of a second fatty streak [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

Main variables in the model are the percentages  $F_k$  of the coronary artery intimal surface area involved with lesions. The index  $k$  indicates the minimal severity of lesions contributing to  $F_k$ . The total percentage of the intimal surface area involved with any type of atherosclerotic lesions is denoted by  $F_1$ . The percentage involved with raised lesions, defined as fibrous plaques or more advanced lesions, is called  $F_2$ . Finally,  $F_3$  relates to the most advanced, complicated lesions. Correspondingly,  $F_0$  refers to the total, healthy or affected, intimal area of the artery wall. Due to the applied normalization it holds  $F_0 = 1$ . Plaque area is a strong predictor of infarction.<sup>13</sup> Complicated lesions are supposed to be most closely related to the risk of wall ruptures and resulting thrombosis and infarction.<sup>14</sup> Risk of myocardial infarction was thus modeled to be proportional to the area involved with complicated lesions  $F_3$ , with a proportionality factor  $v_h$ . A schematic model representation is presented in Figure 1 and an exemplary sequence sketched in Figure 2.

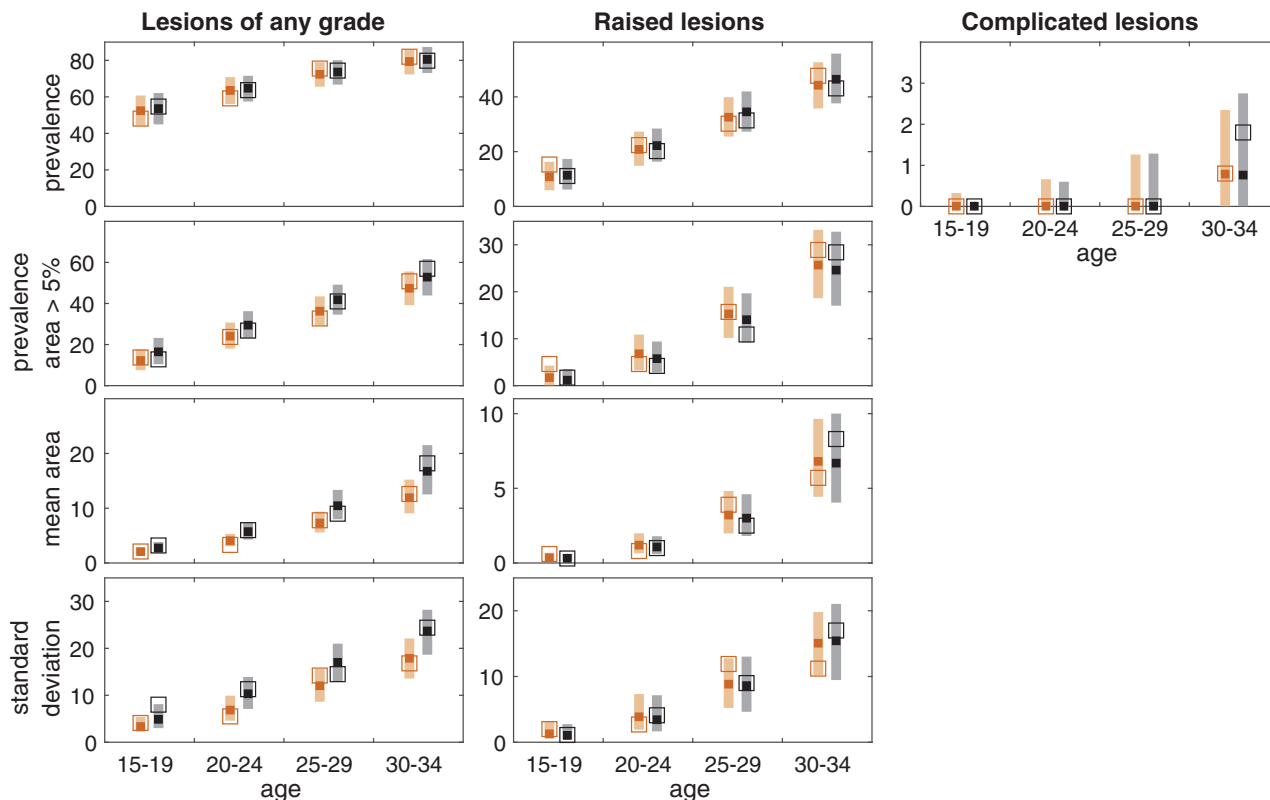
The percentages  $F_k$  are supposed to increase by two processes: formation of new lesions and growth of existing ones. For simplicity, lesion growth is modeled as a deterministic process. As long as the lesion area is small, exponential growth is assumed with a growth rate  $\gamma_k$  (cf. the first term in braces in Equation (1)). On the other hand, initiation of new lesions is assumed to be associated with the permeability or adhesiveness of individual endothelial cells and thus modeled stochastically by a random variable  $T$ , see below. The probability of formation of a new lesion of stage  $k$  is assumed to be proportional to  $F_{k-1}$ . The proportionality factor is called  $v_k$  and the initial area of new lesions is  $s$  (cf. the second term in braces in Equation (1)). When  $F_k$  approaches the area involved with lesions of lower grade  $F_{k-1}$  by either process, growth is saturated (cf. the last term in Equation (1)). To summarize, growth of lesion area in the time span from age  $a$  to age  $a + \Delta a$  is described by the formula:

$$F_k(a + \Delta a) = F_k(a) + \left\{ \gamma_k F_k(a) \Delta a + s T [v_k F_{k-1}(a) \Delta a] \right\} \frac{F_{k-1}(a) - F_k(a)}{F_{k-1}(a)} \quad (1)$$

Here  $T[r]$  denotes a random variable that is close to 1 with probability  $r$  and close to 0 otherwise. Therefore, a new lesion of type  $k$  emerges with probability  $v_k F_{k-1}(a) \Delta a$ . Values between 0 and 1 were allowed in order to avoid discontinuities in the deviance for small changes in the parameters because discontinuities complicate parameter optimization in the fitting procedure. This implicates that new lesions are not strictly generated with probability  $r$  and of size  $s$  but there is also some probability for generation of smaller sized lesions. The random variable  $T[r]$  was implemented by sampling  $x$  uniformly in the interval from 0 to 1 and evaluating

$$t(x, r) = \frac{1}{2} \left( 1 - \tanh \frac{5(x - r)}{r} \right) \quad (2)$$

Bold symbols in Equation (1) denote quantities that vary between simulated individuals. In particular, the inter-individual variability, which is associated with heterogeneity of risk factors in the population, is reflected in the model by a distribution of the growth parameters  $\gamma_k$ . The parameters  $\gamma_k$  are assumed to be normally distributed around  $\gamma_k$  and with relative standard deviation  $\sigma_\gamma$ , that is,  $\gamma_k \sim \mathcal{N}(\gamma_k, \sigma_\gamma \gamma_k)$  but with the constraint that negative values were shifted to zero. No major differences were observed in preliminary analyses whether taking fully positively correlated or uncorrelated  $\gamma_1, \gamma_2$ , and  $\gamma_3$ , and the former was chosen for simplicity. Moreover, the parameters  $v_k$  were tested for a log-normal distribution in preliminary analysis but there was no improvement of fit compared to a model with identical value of  $v_k$  for all simulations. Therefore, there are two sources of heterogeneity for disease development in the model: chance as mediated by the random variable  $T$ , and inter-individual diversity of risk factors as described by the distribution of  $\gamma_k$ .



**FIGURE 3** Simulated and observed age-dependent lesion spread in males. Solid boxes and error bars show the median and 2.5% and 97.5% percentiles of all simulated runs. Orange corresponds to white, black to black men. Open squares represent data derived from 109 to 168 autopsy cases.<sup>12</sup> Shown are the prevalence of lesions (in %), the prevalence of lesions that involve at least 5% of the coronary artery (in %), and the mean and standard deviation of the intimal surface area of the coronary artery involved with lesions (in % of the total intimal surface area) [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

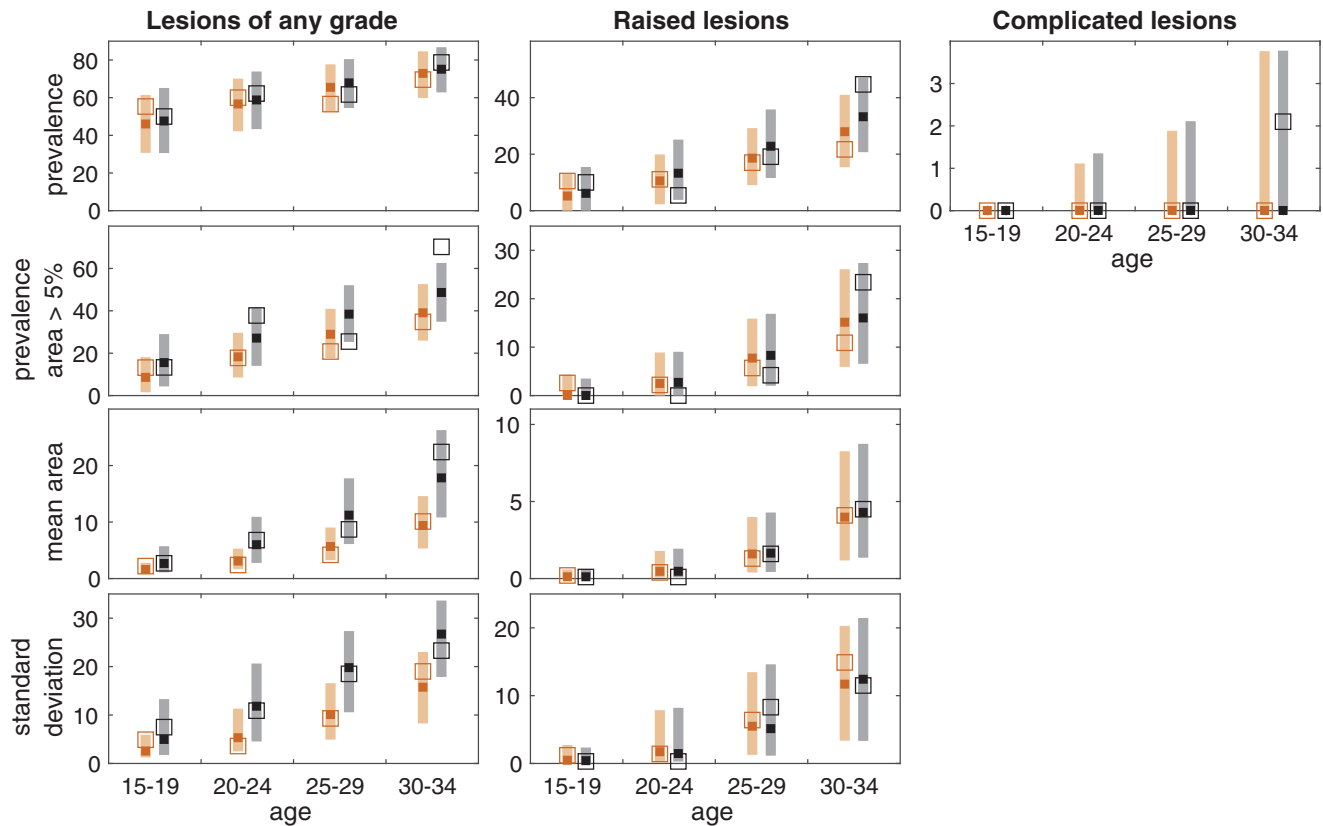
## 2.2 | Data

### 2.2.1 | Subclinical data

Two data sources were used for construction of the model. The first is the large US American pathology study PDAY for which arteries from 1532 persons were collected during the years 1987 to 1990.<sup>12</sup> Arteries were retrieved from persons who had died from external causes between age 15 to 34 years. Thoracic and abdominal aorta, and the right coronary artery were investigated. Motivated by myocardial infarction as the final endpoint in this study, only data for the right coronary artery are considered here. The types of lesions were quantified as well as their extent, measured as the total percentage of intimal surface area involved. Data were evaluated in 5-year age categories separately for male and female sex, and for white and black ethnicity. Prevalence of lesions, prevalence of significant lesions, defined to cover at least 5% of the intimal surface area, mean area (in percent of the total intimal surface area) and their standard deviations were presented for all lesions and for raised lesions. For complicated lesions only the prevalence was presented. These data are displayed as open squares in Figures 3 and 4 together with simulation results as referred to from the Results section.

### 2.2.2 | Epidemiological data

The second data source is the KORA myocardial infarction registry<sup>15,16</sup> which records myocardial infarction within the region of Augsburg, Germany. This study is based on the age interval 25 to 79 and relates to the years 2009 to 2016. On average, 221 745 males and 225 950 females have lived in this region in the relevant age range. Based on accurate case ascertainment, 5310 first incidences of myocardial infarction were registered for males, and 2322 for females. Person-years at risk were estimated from the local population register as outlined in Appendix A.



**FIGURE 4** Simulated and observed age-dependent lesion spread in females. Solid boxes and error bars show the median and 2.5% and 97.5% percentiles of all simulated runs. Orange corresponds to white, black to black women. Open squares represent data derived from 30 to 53 autopsy cases.<sup>12</sup> Shown are the prevalence of lesions (in %), the prevalence of lesions that involve at least 5% of the coronary artery (in %), and the mean and standard deviation of the intimal surface area of the coronary artery involved with lesions (in % of the total intimal surface area) [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

### 2.3 | Model implementation and fitting procedure

The model was implemented in C++ and can be downloaded from.<sup>17</sup> Time steps  $\Delta a$  of one year were applied. Analysis was performed separately for the different sexes. Data from white and black persons were fitted together, with a relative difference in parameters:  $v_k^w = v_k$  for white, and  $v_k^b = v_k \exp(\eta_{v_k})$  for black persons and likewise for  $\sigma_\gamma$ ,  $\gamma_k$ , and  $s$ . The relative difference was constrained not to exceed a factor of two. Beyond that, possible shifts of parameter values during menopause were investigated for females. The age dependence was tested with the functional form  $v_k(a) = v_k[1 + \omega_{v_k} t(a, 50)]$ , and analogous for other parameters, with  $t(a, 50)$  being defined in Equation (2). The function  $t(a, 50)$  is close to 1 at young age and decreases almost to 0 above age 50.

As a measure of goodness-of-fit we apply the deviance. It is given by twice the negative logarithm of the likelihood. To calculate the deviance, 10 000 simulation “runs” were performed. The number of individual model simulations per run was adjusted to the number of observations in the subclinical PDAY data.<sup>12</sup> This led to at least 1 000 000 individual model simulations for the evaluation of the deviance, which took about two seconds on an eight-core desktop computer. The total deviance is given by the sum of the deviance related to the subclinical PDAY and the deviance related to KORA incidence data,<sup>15</sup> see Appendices B and C for details. Therefore, minimizing the total deviance, the model is fitted simultaneously to both data sets. Minimization was performed with MINUIT2 (version 5.28.0). Parameters for ethnicity and menopause were added sequentially, starting with the parameter with highest deviance improvement. For inclusion of parameters and calculation of 95% confidence intervals, the cut point of 3.84 was applied from the  $\chi^2$ -distribution.

### 3 | RESULTS

#### 3.1 | Stability of the fit to simulated data

As model results are obtained by simulation, residual uncertainty remains due to the finite number of simulation runs. The influence of this uncertainty on the total deviance was estimated by repeated model fits. For this purpose, data were applied for males only with no distinction between ethnic groups.

The results from repeated model fitting exhibited a standard deviation of 0.3 in the deviance. Therefore, model selection and parameter estimation are possible accepting some additional uncertainty due to the finite number of simulations.

#### 3.2 | Fit with age-constant parameters

Parameters were added to describe the difference between white and black men in any of the parameters  $\gamma_k, \sigma_\gamma, v_k, s$ . The optimal deviance reduction was obtained for ethnicity-dependent  $\gamma_1$  with a highly statistical improvement of 16.7 ( $p < 0.001$ ). Allowing in addition for ethnicity-dependent  $\gamma_2$  reduced the deviance further by 10.2 ( $p = 0.001$ ), yielding a total deviance of 869.9. The deviances of this stepwise model adjustment can be found in Table 1. Ethnicity-dependent  $v_2$  led to a similar improvement as did  $\gamma_2$  but when taking into account dependence of  $\gamma_2$  on ethnic group, there was no other significant dependence. Maximum likelihood estimates of the parameters and confidence intervals can be found in Table 2. Notably, growth dynamics of fatty streaks was rather low, governed by an increase of  $\gamma_1 = 11\%$  per year. Once a

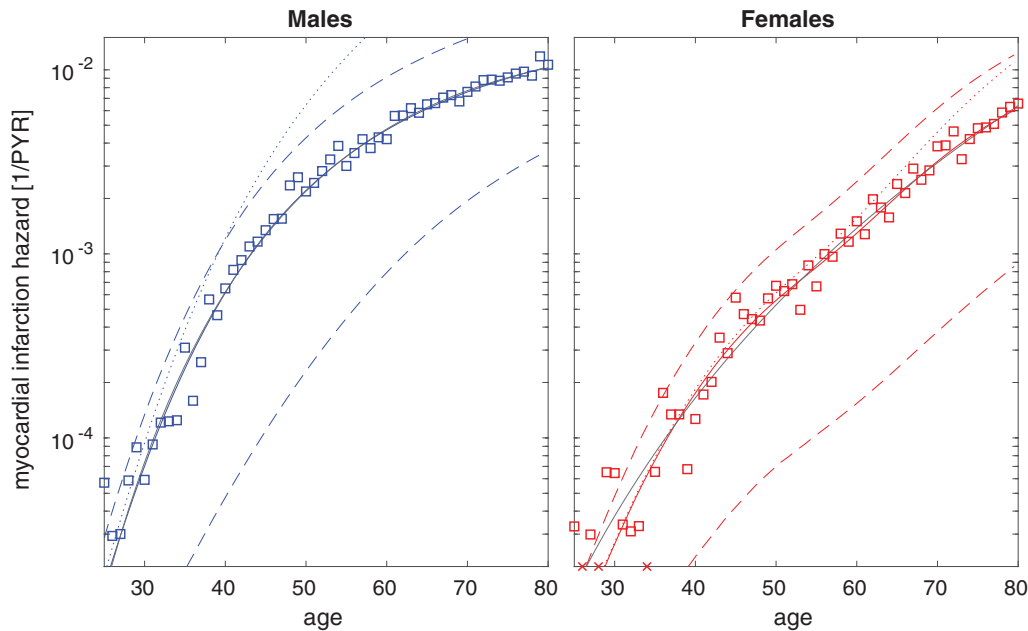
Model description	Males	Females
Initial model	896.8	909.1
+ ethnicity dependent $\gamma_1$	880.1	889.8
+ ethnicity dependent $\gamma_2$	869.9	880.3
+ variation of $\gamma_3$ during menopause		868.6
+ variation of $v_h$ during menopause		859.3

**TABLE 1** Deviance for model variants successively adding new parameters

	Males constant rates	Females constant rates	Females considering menopause
$s$	0.85 (0.67; 1.1)	0.71 (0.43; 1.1)	0.82 (0.48; 1.2)
$\log v_1$	-3.1 (-3.2; -3.0)	-3.4 (-3.5; -3.2)	-3.3 (-3.5; -3.1)
$\gamma_1$	0.11 (0.09; 0.13)	0.13 (0.10; 0.16)	0.11 (0.07; 0.15)
$\sigma_\gamma$	0.56 (0.42; 0.80)	0.42 (0.27; 0.60)	0.60 (0.33; 1.1)
$\log v_2$	0.27 (0.03; 0.50)	-0.33 (-0.79; 0.10)	-0.38 (-0.85; 0.07)
$\gamma_2$	0.35 (0.23; 0.53)	0.47 (0.28; 0.80)	0.37 (0.19; 0.70)
$\log v_3$	-3.9 (-5.2; -3.2)	-2.4 (-3.2; -1.4)	-2.9 (-4.2; -0.93)
$\gamma_3$	0.31 (0.11; NA)	0.05 (0.04; 0.07)	0.09 (0.04; 0.24)
$\log v_h$	-1.0 (-3.6; -0.3)	-1.6 (-2.5; -0.69)	-1.8 (-3.9; 0.20)
$\eta_{\gamma_1}$	0.19 (0.12; 0.26)	0.32 (0.21; 0.45)	0.35 (0.22; 0.49)
$\eta_{\gamma_2}$	-0.53 (-0.69; -0.20)	-0.69 (-0.69; -0.32)	-0.69 (-0.69; -0.22)
$\omega_{\gamma_3}$			-0.63 (-1.0; -0.25)
$\omega_{v_h}$			2.0 (0.27; NA)

**TABLE 2** Maximum likelihood estimates and 95% confidence intervals from the likelihood profile

Rates ( $\gamma_k, v_k, v_h$ ) are given per year, and  $s$  is given as percentage of the coronary artery intimal surface area. Parameters  $\eta_{\gamma_1}, \eta_{\gamma_2}$  denote the logarithm of the relative difference between white and black persons. Parameters  $\omega_{\gamma_3}, \omega_{v_h}$  denote the relative parameter increment for young ages normalized to the time after menopause.



**FIGURE 5** Simulated and fitted myocardial infarction incidence rates. Solid black lines show the best fit with a descriptive model, solid colored lines the simulated mechanistic model (almost masked by the black line for males). Dashed lines illustrate the simulated inter-individual variability. Dotted lines show the mean simulated incidence rate, if the drop out of individuals after myocardial infarction was ignored. Therefore, dotted lines correspond to the average risk expectation of a young individual which is above the age-dependent population mean due to selection effects. For details see the Discussion section. Boxes denote observed rates in the KORA study, and crosses those ages without any observed case during follow up. Data points have been corrected for calendar and birth year dependence [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

**TABLE 3** Deviance related to the fit to the KORA incidence data

	Descriptive model	Mechanistic model constant rates	Mechanistic model considering menopause
Males	453.2	450.2	
Females	461.2	477.6	457.7

raised lesion is formed in the model, it expands much faster, by  $\gamma_2 = 35\%$  per year, and typically dominates after about one decade. Moreover, formation of raised lesions is rather frequent (large  $v_2$ ) but formation of complicated lesions is not (small  $v_3$ ). Parameters differed substantially between black and white men. The growth rate  $\gamma_1$  was larger for black men by a factor  $\exp(0.19) = 1.21$  and was smaller for  $\gamma_2$  by a factor  $\exp(-0.53) = 0.59$ . The modeled course of lesion development is plotted in Figure 3 and the incidence rate in Figure 5.

For females we start again with constant parameters and no distinction between black and white women. Differences between ethnic groups could again be traced back to  $\gamma_1$  and  $\gamma_2$ . Taking into account ethnicity in these parameters led to a 19.4 ( $p < 0.001$ ) and 9.5 ( $p = 0.002$ ) point deviance improvement, yielding finally a total deviance of 880.3. Again, maximum likelihood estimates of the parameters can be found in Table 2. Most maximum likelihood estimates are similar between sexes. Differences exist in the formation rates of advanced lesions: While the early lesions form with lower rates  $v_1, v_2$  in females, the formation rate of complicated lesions  $v_3$  was higher for females compared to males. Another important difference lies in the growth rate of complicated lesions  $\gamma_3$  which was much smaller for females.

### 3.3 | Comparison to a descriptive model

A peculiar property of the mechanistic model is its ability to fit simultaneously data on lesion development and myocardial infarction incidence. In fact, it is not meaningful to fit the lesion data without a model that relates the different stages. The incidence data, however, can also be fitted with standard descriptive parametric models. Applying the descriptive model set out in Appendix A, a deviance of 453.2 is achieved for males and 461.2 for females.

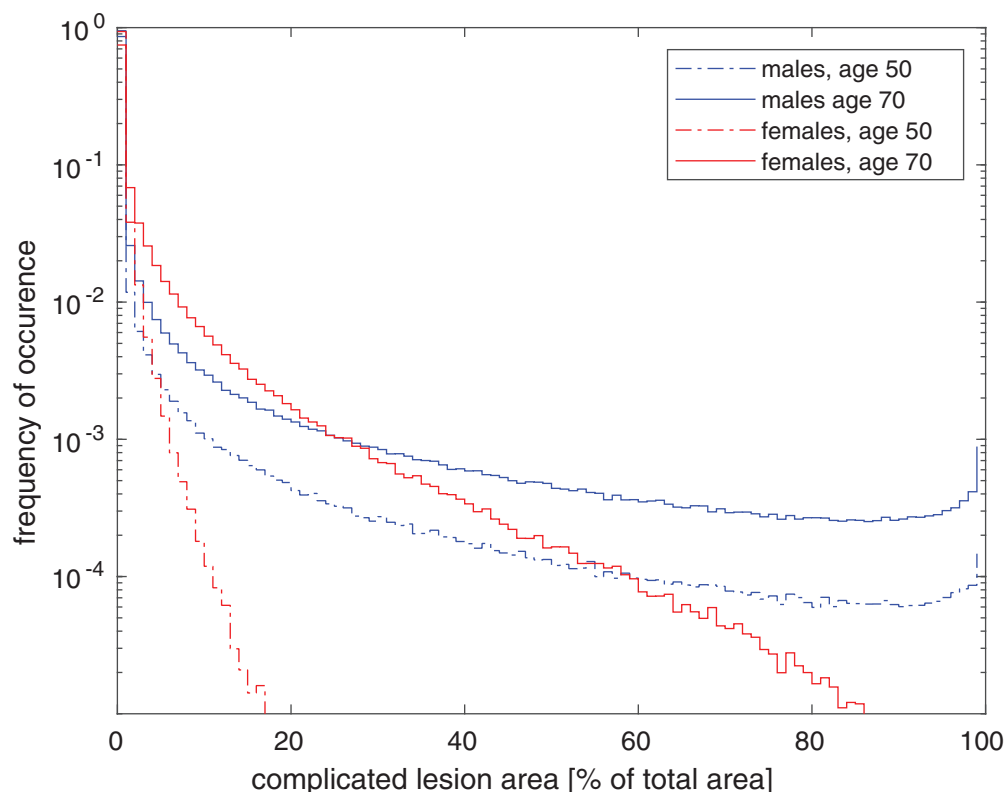
The mechanistic model with constant parameters yielded a total deviance of 869.9 for males and 880.3 for females. For comparison, only the part of the total deviance related to the fit to the KORA incidence data is relevant. It is shown in Table 3. For males it evaluated to 450.2, somewhat superior to the descriptive model. On the other hand, it evaluated to 477.6 for females, thus not reaching the same goodness-of-fit as the descriptive model. In this regard, it should be noted that the only purpose of the descriptive model is to yield a good fit to the incidence data, and only three parameters were necessary to achieve this goal. In contrast, the mechanistic model simultaneously fits other data and was compiled to describe a disease evolution. Therefore, it is necessarily much more complex (nine parameters relevant to white women) and this difference may easily account for a few points in the deviance difference in either direction. However, a deviance difference of 16.4 points as observed for females indicates some process not being well described by the mechanistic model. Indeed, constant parameters were used so far for both sexes despite the well-known protective hormonal effects before menopause.

### 3.4 | Including age dependencies due to menopause

Possible shifts of the parameters  $v_k$ ,  $v_h$ , and  $\gamma_k$  during menopause were investigated. Notation was chosen such that  $v_k$ ,  $v_h$ ,  $\gamma_k$  denote the parameter values after menopause while values before menopause were denoted by products  $v_k(1 + \omega_{v_k})$  etc.

Significant improvements in the deviance were observed only for  $\gamma_3$  and  $v_h$ , that is, for the late stages of disease development. An improvement of 11.7 ( $p < 0.001$ ) was achieved when  $\omega_{\gamma_3} = -1$ . Even lower values would mean a general regression of complicated lesions before the age of menopause and were not allowed in the model fit. For  $v_h$ , the deviance improved by 10.9 ( $p = 0.001$ ) points with a best fit value  $\omega_{v_h} = 1.1$ . When allowing both parameters to be age dependent, effect size was strengthened in  $v_h$  and weakened in  $\gamma_3$ . The combined deviance improvement was 21.0 points, yielding a total deviance of 859.3, see Table 1. The part of the deviance related to the fit to the KORA incidence data evaluated to 457.7 which is 3.5 points lower than the one from the descriptive model.

Maximum likelihood estimates and confidence intervals can be found in Table 2. Compared to the fit with constant parameters, maximum likelihood estimates for  $v_3$  and  $\gamma_3$  somewhat converged to the values for males though  $\gamma_3$  remained to be significantly smaller for females. The modeled course of lesion development is plotted in Figure 4 and the incidence rate in Figure 5. Despite rather similar parameter values, there are striking differences in incidence curves between males and females: A strong increase in risk until middle age followed by a pronounced flattening is only observed in men.



**FIGURE 6** Simulated size distribution of complicated lesions in the population for age 50 and age 70 as percentage of the coronary artery intimal surface area [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



The flattening is related to the size distribution of complicated lesions as explained in more detail in the Discussion section. The simulated size distribution of complicated lesions is shown in Figure 6. For the examples presented, the majority of simulated persons do not show any complicated lesion. However, for females the frequency of occurrence drops much faster with the size of lesions than for males. For more than 0.1% of (simulated) males at age 50, complicated lesions involve areas larger than 80% of the intimal surface area. On the other hand, for females this fraction is much lower even at age 70 although myocardial infarction risk is higher for females at age 70 compared to males at age 50 (see Figure 5).

## 4 | DISCUSSION

The aim of mechanistic modeling of epidemiological data is to relate the age-incidence curve to subclinical disease development. This involves both a top down approach, as inference is made from the incidence curve about the dynamics of disease development, and a bottom up approach, as information on biological processes is taken into account for risk estimation. For the first time, a mechanistic model was built from a combined fit to subclinical and incidence data. Several stages of subclinical evolution were modeled. Not only the average evolution but also the variation in the study cohort were reproduced very well (see Figures 3 to 5 and Table 3). This was achieved by appropriate complexity of the model, which was solved by simulation. Nevertheless, stable maxima of the likelihood could be found with similar parameter best estimates between genders (Table 2), demonstrating the technical feasibility of the approach.

Properties of the resulting model parameters can be compared to observation. For example, the growth rate of early lesions is higher for black ethnicity in our model, see Table 2. On the other hand, black ethnicity is associated with a lower growth rate of raised lesions. This is consistent with results of an ultrasound study, part of the Multi-Ethnic Study of Atherosclerosis: black ethnicity was positively related to carotid intima media thickness, which is an imaging marker of early atherosclerosis, and it was inversely related to the formation of plaques.<sup>18</sup> Gender differences can be discussed accordingly. In the model, rates of formation and growth of early lesions are very similar for the different genders, see Table 2. Formation of raised lesions, however, occurs at a lower rate for females. Indeed, lesions in women exhibit less lipid cores, the main characteristic of raised lesions.<sup>19-22</sup> Finally, while formation of complicated lesions may be at a higher rate in the model for females, their growth is markedly slower. This may be understood by different complicated lesion phenotypes prevailing in males and females: Lesions in women appear to be more stable, they show less ruptures but more erosions.<sup>19-21</sup>

To achieve a good fit for females, it was necessary to allow for age dependence in model parameters. This, however, was to be expected as cardiovascular function is altered in females around the age of menopause,<sup>23,24</sup> probably caused by higher female hormone levels before menopause. In the preferred fit two parameters were age dependent: The growth rate of complicated lesions was reduced before menopause but the risk of rupture of established complicated lesions enhanced. Indeed, increasing female hormone levels by postmenopausal hormone replacement therapy has been observed to inhibit atherosclerosis progression but to promote plaque instability.<sup>25</sup> These opposing mechanisms make plausible why hormone replacement therapy is most effective when started immediately after menopause:<sup>25</sup> at this age there are essentially no plaques yet that could become unstable. Notably, the two opposing effects were revealed by the model fit despite the fact that subclinical data were available only up to age 34. Therefore, this can be viewed as an example for the incidence curve harboring information on disease dynamics which can be unraveled with mechanistic modeling.

After age at menopause, the incidence curve for females is close to an exponential increase (see right panel of Figure 5). In contrast, for males there is a sharp increase of risk at younger ages followed by a pronounced flattening that finally yields little sex difference in the risk for very old ages. This disparity can be observed also in other data.<sup>26</sup> In the model, the flattening is explained by two mechanisms. The first is saturation of advanced lesion growth. The second mechanism relates to what is usually called frailty in the statistical literature<sup>27</sup> and occurs generically also for other multi-stage models.<sup>28</sup> Those individuals with highest risk are most likely to experience an infarction and to be thus excluded from the incidence cohort. This effect reduces the average risk in the remaining cohort. Mathematically, this effect is taken into account by the exponential terms in Equation (B1) in Appendix B. Ignoring this effect, thus assuming the incidence rate  $h$  to be given just by  $v_h F_3$ , would yield considerably higher values, as depicted in Figure 5 by the dotted lines. Discrimination between the different mechanisms has practical implications: Individual risks do not follow the population hazard. Saturation of advanced lesion growth implies that the risk increase is attenuated in persons with almost maximal lesions. In contrast to that, frailty effects imply that individual risks may increase without any damping even if there is a flattening of the population hazard. While frailty is always there, it would require a huge spread in individual risks

if it was the only mechanism leading to the flattening. Although there is considerable variation in individual risk in our model (see Figure 6 noting that in the model risk is proportional to the intimal surface area involved with complicated lesions), saturation, too, plays a role in the flattening. In the model, variation in risk follows from the observed extent of lesion variation. Therefore our results add to the evidence that despite the relevance of frailty on the simulated population hazard, there is still also flattening of individual risks.

Variation in risk is illustrated in Figure 5 by the dashed lines. The upper (lower) dashed line corresponds to the hazard of the more (less) predisposed half of the population. In the model this corresponds to individual lesion growth rates above (below) the population median, defined at young age. Interestingly, only up to an age of about 50 years variability increases with age. At higher ages, simulated variability remains rather constant for females and decreases for males. In real life cohorts, variability in risk may change due to age-dependent prevalence and biological effectiveness of risk factors. These may be the main reasons why at advanced age the effect size of risk factors tends to decrease.<sup>26,29</sup> However, our model results show that some damping in the divergence of individual risks can be expected even with permanent risk factors.

Obviously our simple model can not capture any detail. First, a single risk factor typically acts only on some stage of disease development, which may lead to a peculiar age- and time dependent risk. In contrast, the simulated variability is based on a single, overall measure of individual susceptibility that affects all stages of disease development. Second, while our analysis shows that several dynamical features can be explained even without or with very simplistic age dependence of the parameters, this does not preclude the existence of biological alterations of the development of atherosclerosis with age. More generally, the necessary simplification of the true disease development is an inevitable limitation of any mechanistic study. Especially relevant points may include here that spatial distribution of lesions is ignored, lesions are characterized only by few stages, dynamics are limited to a simple saturating growth model, and that there is no feedback from late-stage to early-stage lesions. Finally, model results in this study are subject to some numerical uncertainty due to limitations in the number of simulations.

Despite these limitations, some relations of age incidence patterns and underlying mechanisms could be discussed with the model. In the future the model may be extended to describe the dynamics of risk after short-term or protracted exposure to risk factors. This is in analogy to cancer for which mechanistic models have been applied to many cohorts.<sup>4</sup> However, it also goes well beyond that. Importantly, cardiovascular disease development is directly accessible to observation. In this study, this has been exploited by taking into account subclinical data from an autopsy study. The autopsy study obtained a level of detail that is not achievable with imaging, thus allowing to model early lesion development. On the other hand, longitudinal information could clearly advance modeling of lesion growth and rupture. Indeed, a first step in this direction is planned: to model personalized cardiovascular risks from radiotherapy in females with breast cancer. This study group is special as radiation treatment poses a well quantifiable risk factor and as the extent of atherosclerotic lesions can be inferred from the treatment planning CTs. Another interesting point in cardiovascular disease modeling is to study not only the effect of risk factors but also of various treatment measures, to improve the assessment of individual long-term benefits. Finally, it should be noted, that the approach of simulating epidemiological data based on subclinical pre-stages is a very general approach that may provide a new tool for understanding of long-term dynamics of various diseases.

## 5 | CONCLUSION

Mechanistic modeling of epidemiological data allows to draw connections between features in the age-incidence curve and the biological disease development. In order to model the development of atherosclerosis and subsequent myocardial infarction, we applied a simulation approach for the first time. Waiving analytical solvability allows for more complex models and for taking into account simultaneously randomness and individual variability in the disease development. Moreover, in order to ensure a realistic model behavior for early, subclinical disease development, we fitted the model not only to epidemiological but also to pathological data. This new approach can readily be extended to explore the implications of more detailed biological processes, or can be applied to other diseases. In particular, fascinating options are the inclusion of imaging data and personalized model predictions.

## ACKNOWLEDGMENTS

Special thanks go to Noemi Castelletti for fruitful discussions, proofreading of the manuscript, and help in the preparation of graphics. The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research

Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. This project has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 755523 (MEDIRAD). Open access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST

The authors declare no potential conflict of interests.


## AUTHOR CONTRIBUTIONS

CS and JCK conceived of the presented study. CS developed the software, designed and performed the analysis, and wrote the paper with input from JCK and SR. SR, MH, CM, AP contributed data. All authors discussed the results and reviewed the manuscript.

## DATA AVAILABILITY STATEMENT

Software and data that support the findings of this study are openly available in figshare at <https://doi.org/10.6084/m9.figshare.14572869.v1>.

## ORCID

Cristoforo Simonetto  <https://orcid.org/0000-0003-4816-3514>

## REFERENCES

- Hougaard P. Multi-state models: a review. *Lifetime Data Anal.* 1999;5(3):239-264. <https://doi.org/10.1023/a:1009672031531>.
- Andersen PK, Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res.* 2002;11(2):91-115. <https://doi.org/10.1191/0962280202SM276ra>.
- Peto R. Epidemiology, multistage models, and short-term mutagenicity tests. *Int J Epidemiol.* 2016;45(3):621-637. <https://doi.org/10.1093/ije/dyv199>.
- Rühm W, Eidemüller M, Kaiser JC. Biologically-based mechanistic models of radiation-related carcinogenesis applied to epidemiological data. *Int J Radiat Biol.* 2017;93(10):1093-1117. <https://doi.org/10.1080/09553002.2017.1310405>.
- Moolgavkar SH, Knudson AG. Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst.* 1981;66(6):1037-1052. <https://doi.org/10.1093/jnci/66.6.1037>.
- Little MP, Vineis P, Li G. A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data. *J Theor Biol.* 2008;254(2):229-238. <https://doi.org/10.1016/j.jtbi.2008.05.027>.
- Dewanji A, Jeon J, Meza R, Luebeck EG. Number and size distribution of colorectal adenomas under the multistage clonal expansion model of cancer. *PLoS Comput Biol.* 2011;7(10):e1002213. <https://doi.org/10.1371/journal.pcbi.1002213>.
- Little MP, Gola A, Tzoulaki I. A model of cardiovascular disease giving a plausible mechanism for the effect of fractionated low-dose ionizing radiation exposure. *PLoS Comput Biol.* 2009;5(10):e1000539. <https://doi.org/10.1371/journal.pcbi.1000539>.
- Simonetto C, Azizova TV, Barjaktarovic Z, et al. A mechanistic model for atherosclerosis and its application to the cohort of Mayak workers. *PLoS One.* 2017;12(4):e0175386. <https://doi.org/10.1371/journal.pone.0175386>.
- Lusis AJ. Atherosclerosis. *Nature.* 2000;407(6801):233-241. <https://doi.org/10.1038/35025203>.
- Stary HC, Chandler AB, Glagov S, et al. A definition of initial, fatty streak, and intermediate lesions of atherosclerosis. a report from the committee on vascular lesions of the council on arteriosclerosis, American heart association. *Arterioscler Thromb.* 1994;14(5):840-856.
- Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Research Group. Natural history of aortic and coronary atherosclerotic lesions in youth. findings from the PDAY study. *Arterioscler Thromb.* 1993;13(9):1291-1298.
- Spence JD, Eliasziw M, DiCicco M, Hackam DG, Galil R, Lohmann T. Carotid plaque area: a tool for targeting and evaluating vascular preventive therapy. *Stroke.* 2002;33(12):2916-2922. <https://doi.org/10.1161/01.str.0000042207.16156.b9>.
- Stary HC, Chandler AB, Dinsmore RE, et al. A definition of advanced types of atherosclerotic lesions and a histological classification of atherosclerosis. a report from the committee on vascular lesions of the council on arteriosclerosis, American heart association. *Circulation.* 1995;92(5):1355-1374. <https://doi.org/10.1161/01.cir.92.5.1355>.
- Löwel H, Meisinger C, Heier M, Hörmann A. The population-based acute myocardial infarction (AMI) registry of the MONICA/KORA study region of Augsburg. *Gesundheitswesen.* 2005;67(Suppl 1):31-37. <https://doi.org/10.1055/s-2005-858241>.
- Kuch B, Heier M, Von Scheidt W, Kling B, Hoermann A, Meisinger C. 20-year trends in clinical characteristics, therapy and short-term prognosis in acute myocardial infarction according to presenting electrocardiogram: the MONICA/KORA AMI registry (1985-2004). *J Intern Med.* 2008;264(3):254-264. <https://doi.org/10.1111/j.1365-2796.2008.01956.x>.

17. Simonetto C. athsim: Simulate atherosclerosis and subsequent myocardial infarction. figshare, Software 2021. <https://doi.org/10.6084/m9.figshare.14572869.v1>.
18. Tattersall MC, Gassett A, Korcarz CE, et al. Predictors of carotid thickness and plaque progression during a decade: the multi-ethnic study of atherosclerosis. *Stroke*. 2014;45(11):3257-3262. <https://doi.org/10.1161/STROKEAHA.114.005669>.
19. Lansky AJ, Ng VG, Maehara A, et al. Gender and the extent of coronary atherosclerosis, plaque composition, and clinical outcomes in acute coronary syndromes. *JACC Cardiovasc Imaging*. 2012;5(Suppl 3):S62-S72. <https://doi.org/10.1016/j.jcmg.2012.02.003>.
20. Sangiorgi G, Roversi S, Zoccai GB, et al. Sex-related differences in carotid plaque features and inflammation. *J Vasc Surg*. 2013;57(2):338-344. <https://doi.org/10.1016/j.jvs.2012.07.052>.
21. Kataoka Y, Puri R, Hammadah M, et al. Sex differences in nonculprit coronary plaque microstructures on frequency-domain optical coherence tomography in acute coronary syndromes and stable coronary artery disease. *Circ Cardiovasc Imaging*. 2016;9(8):e004506. <https://doi.org/10.1161/CIRCIMAGING.116.004506>.
22. Haaf MET, Rijndertse M, Cheng JM, et al. Sex differences in plaque characteristics by intravascular imaging in patients with coronary artery disease. *EuroIntervention*. 2017;13(3):320-328. <https://doi.org/10.4244/EIJ-D-16-00361>.
23. Witteman JC, Grobbee DE, Kok FJ, Hofman A, Valkenburg HA. Increased risk of atherosclerosis in women after the menopause. *BMJ*. 1989;298(6674):642-644. <https://doi.org/10.1136/bmj.298.6674.642>.
24. Hayward CS, Kelly RP, Collins P. The roles of gender, the menopause and hormone replacement on cardiovascular function. *Cardiovasc Res*. 2000;46(1):28-49. [https://doi.org/10.1016/s0008-6363\(00\)00005-5](https://doi.org/10.1016/s0008-6363(00)00005-5).
25. Gungor F, Kalelioglu I, Turfanda A. Vascular effects of estrogen and progestins and risk of coronary artery disease: importance of timing of estrogen treatment. *Angiology*. 2009;60(3):308-317. <https://doi.org/10.1177/0003319708318377>.
26. Anand SS, Islam S, Rosengren A, et al. Risk factors for myocardial infarction in women and men: insights from the INTERHEART study. *Eur Heart J*. 2008;29(7):932-940. <https://doi.org/10.1093/eurheartj/ehn018>.
27. Aalen OO. Effects of frailty in survival analysis. *Stat Methods Med Res*. 1994;3:227-243.
28. Moolgavkar SH, Dewanji A, Venzon DJ. A stochastic two-stage model for cancer risk assessment. I. the hazard function and the probability of tumor. *Risk Anal*. 1988;8(3):383-392.
29. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2014;63(25 Pt B):2935-2959. <https://doi.org/10.1016/j.jacc.2013.11.005>.
30. Statistisches Bundesamt (German Federal Statistical Office). Sterbetafel (Periodensterbetafel): Deutschland, Jahre, Geschlecht, Vollen-detes Alter; 2009. [https://www-genesis.destatis.de/genesis/online/link/tabellen/12621\\*](https://www-genesis.destatis.de/genesis/online/link/tabellen/12621*) [Online. Accessed August 6, 2019].

**How to cite this article:** Simonetto C, Rospleszcz S, Heier M, Meisinger C, Peters A, Kaiser JC. Simulating the dynamics of atherosclerosis to the incidence of myocardial infarction, applied to the KORA population. *Statistics in Medicine*. 2021;40:3299–3312. <https://doi.org/10.1002/sim.8951>

## APPENDIX A. DETAILS TO THE FIT TO POPULATION INCIDENCE DATA

Two technical issues arose when fitting to the KORA incidence data.<sup>15</sup> As the data were obtained from a population register, only the number of cases and person years within the study region are known. However, to obtain the person years under risk for first incidence, the share of persons with previous myocardial infarction is also necessary. Therefore we estimated this number using calendar-year averaged myocardial infarction incidence and death rates. At age 25, it was assumed that there is no person with prevalent myocardial infarction. For each further year of age, the estimated number of persons with prior myocardial infarction was obtained by first adding the corresponding myocardial infarction incidence rate, second subtracting the corresponding mortality rate, and, third, relative reduction by general survival. The total survival is not registered in the KORA cohort and German survival rates<sup>30</sup> were used instead. Variation of this rate had practically no impact on our results.

Another issue was the existence of trends in calendar year and birth year. As those trends are not in the focus of the present study, they were accounted for by correction factors that were derived by fitting of a descriptive model. The following mathematical function was applied to fit the age dependence of the incidence hazard:

$$\exp [\beta_0 + \beta_1 \log(a) + \beta_2 \log^2(a)] \quad (\text{A1})$$

Calendar and birth year dependence were best described by the factors

$$\exp [\beta_y(y - 2010)] \times \exp [\beta_b(1940 + a - y)\Theta(1940 + a - y)] \quad (\text{A2})$$

Here  $y$  denotes calendar year and the function  $\Theta$  equals 1 for birth years  $y - a$  before 1940 and 0 otherwise. Best estimates for  $\beta_y$  and  $\beta_b$  were  $-0.010$  and  $0.040$  for males and  $-0.021$  and  $0.050$  for females. The same factor, Equation (A2) was also applied to the mechanistic model and used to correct the data points in Figure 5. Re-fitting of the parameters  $\beta_y$ ,  $\beta_b$  with the mechanistic model had practically no impact on the final results.

## APPENDIX B. THE LIKELIHOOD RELATED TO POPULATION INCIDENCE DATA

The likelihood related to the KORA incidence data<sup>15,16</sup> is given by the product of the likelihoods for each cell of the data characterized by sex, age, and calendar year. For each cell the standard Poisson likelihood was applied, given by  $e^{n-\lambda} (\lambda/n)^n$  with  $n$  the number of observed myocardial infarctions in the data cell and  $\lambda$  the number expected from the model. The number of expected infarctions is the product of person years under risk for the data cell, and the respective incidence rate  $h$  obtained from the simulation. A direct estimate of myocardial infarction rates could be obtained by simulating infarction incidence with rate  $v_h F_3$ , see Figure 1. Then the incidence rate was obtained by the number of simulated infarctions at some age  $a$ , divided by the number of simulated persons without infarction prior to age  $a$ . An equivalent but numerically more stable solution is given by:

$$h(a) = \frac{\sum [v_h F_3 \exp(-\int_0^a v_h F_3 dt)]}{\sum \left[ \exp\left(-\int_0^a v_h F_3 dt\right) \right]} \quad (\text{B1})$$

Here the sums run over all 10 000 times 159 individual model simulations for males and 10 000 times 53 for females, see below. The exponential factor in the nominator equals the probability of a (simulated) person not yet to have suffered from myocardial infarction, that is, still to be at risk at age  $a$ . Analogously, the denominator corresponds to the number of simulated persons without myocardial infarction prior to age  $a$ .

## APPENDIX C. THE LIKELIHOOD RELATED TO SUBCLINICAL DATA

In each run of the simulation, the lesion development is simulated for 159 white and 168 black males or for 53 white and 47 black females. This corresponds each to the maximal number of observations in a single age category of the PDAY data.<sup>12</sup> The same measures of lesion spread were calculated from each simulation run that were also presented in the data: prevalence, prevalence of significant lesions, mean lesion area, and standard deviation of the lesion area. By matching the number of individual model simulations used for these calculations to the respective number of observations in the data the statistical uncertainty is the same in a single simulation run and in the data.

Distributions of the simulated results were obtained by performing 10 000 runs. The likelihood of the simulated model given the data is taken as the probability of the data under these distributions. These distributions were approximated by log-normal distributions based on the respective geometric mean and standard deviation of the 10 000 results. The geometric mean is real valued only on a set of positive numbers. Therefore, if for any run a measure of lesion spread was below 0.05, it was substituted by 0.05, half the accuracy of the results stated in the data.<sup>12</sup>

The log-normal distribution is not a good approximation close to zero. Therefore, another approach was chosen for prevalence data with less than three cases in the data. To avoid discontinuities, we defined the individual probability  $p_i$  for the existence of a lesion of type  $k$  by the smooth function  $1 - t(\mathbf{F}_k, s)$  and, accordingly, for the existence of a significant lesion by  $1 - t(\mathbf{F}_k, 5\%)$ . The function  $t$  is defined in Equation (2) in the main text. The probability of a run to conform to the data is then  $\prod_i (1 - p_i)$  if there is no lesion observed in the data,  $\sum_j \prod_{i \neq j} p_j (1 - p_i)$  for one person with a lesion, and  $\sum_k \sum_{j \neq k} \prod_{i \neq j, i \neq k} p_k p_j (1 - p_i)$  for exactly two persons

with a lesion. Here indices  $i$ ,  $j$ ,  $k$  run over individual model simulations within a run. The likelihood of the model regarding this prevalence data point was evaluated as the mean of the above probabilities over all 10 000 runs.

Finally, the likelihood related to all subclinical PDAY data is the product of the likelihoods for each calculated measure of lesion spread, lesion type, and category in ethnic group and age. (“One likelihood for each box in Figures 3 and 4 of the main text.”) Normalization of the likelihood is arbitrary.