OPEN ACCESS Freely available online

PLOS ONE

Lexis Diagram and Illness-Death Model: Simulating Populations in Chronic Disease Epidemiology

Ralph Brinks[1,2]*, Sandra Landwehr[1,3], Rebecca Fischer-Betz[2], Matthias Schneider[2], Guido Giani[3]

1 German Diabetes Center, Institute of Biometry and Epidemiology, Duesseldorf, Germany, **2** University Hospital, Polyclinics for Rheumatology, Duesseldorf, Germany, **3** Heinrich-Heine-University, Institute for Statistics in Medicine, Duesseldorf, Germany

Abstract

Chronic diseases impose a tremendous global health problem of the 21st century. Epidemiological and public health models help to gain insight into the distribution and burden of chronic diseases. Moreover, the models may help to plan appropriate interventions against risk factors. To provide accurate results, models often need to take into account three different time-scales: calendar time, age, and duration since the onset of the disease. Incidence and mortality often change with age and calendar time. In many diseases such as, for example, diabetes and dementia, the mortality of the diseased persons additionally depends on the duration of the disease. The aim of this work is to describe an algorithm and a flexible software framework for the simulation of populations moving in an illness-death model that describes the epidemiology of a chronic disease in the face of the different times-scales. We set up a discrete event simulation in continuous time involving competing risks using the freely available statistical software R. Relevant events are birth, the onset (or diagnosis) of the disease and death with or without the disease. The Lexis diagram keeps track of the different time-scales. Input data are birth rates, incidence and mortality rates, which can be given as numerical values on a grid. The algorithm manages the complex interplay between the rates and the different time-scales. As a result, for each subject in the simulated population, the algorithm provides the calendar time of birth, the age of onset of the disease (if the subject contracts the disease) and the age at death. By this means, the impact of interventions may be estimated and compared.

Citation: Brinks R, Landwehr S, Fischer-Betz R, Schneider M, Giani G (2014) Lexis Diagram and Illness-Death Model: Simulating Populations in Chronic Disease Epidemiology. PLoS ONE 9(9): e106043. doi:10.1371/journal.pone.0106043

Editor: Catalin Buiu, Politehnica University of Bucharest, Romania

Received May 21, 2014; **Accepted** July 27, 2014; **Published** September 12, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: Financial support for this study was provided by a grant from the Hiller Foundation, Erkrath, Germany. The funder had no role in study design, data collection, analysis and interpretation, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: ralph.brinks@ddz.uni-duesseldorf.de

Introduction

Chronic diseases impose a tremendous global health problem of the 21st century. The World Health Organization estimates that 63% of all deaths in 2008 were caused by chronic diseases [1]. Besides taking measures in politics and society, research efforts are needed to oppose this threat. In studying the characteristics of chronic diseases from a public health perspective, it is often important to consider different time-scales [2]. The age of the subjects in a population is a risk factor for many diseases. Changes in life-style and medical care influence the risk of contracting and dying from the disease on a secular scale. Thus, the incidence rate of the chronic disease as well as the mortality rates (with and without the disease) depend on the calendar time. Moreover, the mortality of people with the disease often depends on the duration since its onset. Examples are diabetes [3,4], dementia [5], depression [6], and systemic lupus erythematosus [7]. For decision-makers all time-scales may be important.

As a hypothetical example, consider the question which health programme to choose from two possibilities A and B if the outcome of interest is the gain of life-years. Possibility A is known to decrease the incidence of the disease by 15%, and possibility B lowers the mortality of those having the disease for more than ten years by 50%. The decision depends on several factors. If, on the one hand, the incidence rate is non-zero for children only and the birth rate in the population is low, possibility A may have little effect with respect to the gain of life-years. If, on the other hand, the chronic disease has very few people reaching ten-year survival after onset, programme B can be nearly useless. In problems similar to the example, the decision-maker may face a complex interplay of epidemiological and demographical considerations.

The aim of this work is to describe an algorithm and a flexible software framework for simulation of populations moving in a multi-state model (illness-death model) that describes a chronic disease. The simulation takes into account the different time-scales: calendar time, age, and duration of the disease. Although simulations using multi-state models are subject to recent textbooks [8], to our knowledge no algorithm has been described that incorporates the effects of all the different time-scales.

Methods

A popular framework for studying irreversible diseases is the illness-death model (IDM) consisting of the three states *Normal*, *Disease* and *Death*, [9–11]. The associated transition rates, synonymously *densities* (in units "per person-time", not to be confused with risks or probabilities [12]), are the incidence i, and the mortality rates m_0 and m_1 (Figure 1). In general, these rates

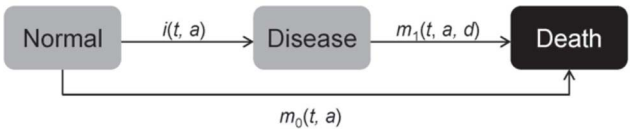

Figure 1. Three states model of normal (healthy), diseased and dead subjects. The transition rates may depend on calender time t, age a, and in case of m_1 also on the duration d of the disease.
doi:10.1371/journal.pone.0106043.g001

depend on calendar time (t), age (a) and in case of m_1 on the duration of the disease (d).

This article presents a method for simulating populations moving in the IDM. The motivation for the algorithms comes from analytical epidemiology where relations between common epidemiological measures are studied. Examples for those measures are the prevalence, the duration of a disease, the age of onset (or diagnosis), and lost life-years (due to the disease). A typical question may be: what is the mean age of diagnosis of subjects born in a certain time period? What is their mean age at death? Another interesting aim is the estimation of the incidence rate i from cross-sectional information. At a specific point in time t', each of the subjects $j=1,\ldots,n$, has a unique "status". Neglecting those who are unborn or dead at t', the status in the IDM is either *normal* (non-diseased) or *diseased*. Thus, the status can be seen as a binary random variable, and data of this kind are typically called *current status data* [13]. The current status is closely linked with the incidence i and the mortalities m_0 and m_1 before t'. Estimating the incidence from current status data, for example, has been a topic in research for decades [14]. The framework presented here may be useful in this field.

Overview of the simulation algorithm

The simulation is a microsimulation, i.e., it treats each person in the population as an autonomous unit. For each person, indexed $j, j=1,\ldots,n$, the relevant events *diagnosis* and *death* are simulated. This is accomplished in two steps:

1. Contracting the disease or dying without the disease is modelled as competing risk [11]. Given the time $t_0^{(j)}$ of birth of person j, the cumulative distribution function $F_1^{(j)}$ of the *first failure time* $T_1^{(j)}$ is

$$F_1^{(j)}(t)=1-\exp\left(-\int_0^t i\left(t_0^{(j)}+\tau,\tau\right)+m_0\left(t_0^{(j)}+\tau,\tau\right)\mathrm{d}\tau\right). \quad (1)$$

   The term *first failure time* $T_1^{(j)}$ refers to the time of diagnosis or death without disease and is measured in time units after birth of person j. Thus, $T_1^{(j)}$ is the age at which the first transition from the state *Normal* occurs. Given that a transition occurs at $T_1^{(j)}$ for person j, then the odds of moving into state *Disease* versus moving into state *Death* is

$$\frac{i\left(t_0^{(j)}+T_1^{(j)},T_1^{(j)}\right)}{m_0\left(t_0^{(j)}+T_1^{(j)},T_1^{(j)}\right)}.$$

2. If the event at $T_1^{(j)}$ is the death (without the disease), the simulation for person j is finished. If, however, the event is the diagnosis of the disease, the "second failure time" $T_2^{(j)}$ to death (with disease) has the distribution function $F_2^{(j)}$:

$$F_2^{(j)}\left(t|T_1^{(j)}\right)=1-\exp\left(-\int_0^t m_1\left(t_0^{(j)}+T_1^{(j)}+\tau,T_1^{(j)}+\tau,\tau\right)\mathrm{d}\tau\right). \quad (2)$$

The next section describes in detail how the integrals in Equations (1) and (2) may be calculated in the simulation. After calculating the integrals, the question arises how the times T_1 and T_2 can be obtained from F_1 and F_2. This is done by the *inverse transform sampling method*: Let F be a cumulative distribution function and $u\in(0,1)$. For $F^{-1}(u) := \inf\{x|F(x)\geq u\}$ it holds: If U is a uniform random variable on (0,1), then $F^{-1}(U)$ follows the distribution F. Thus, the simulation of T_1 and T_2 is easy, if a random number generator for U such as `runif` in R is available.

For each of the n persons in the population we store four pieces of data:

1. a unique identifier j,
2. the date $t_0^{(j)}$ of birth (`dob`) of person j,
3. the age at diagnosis (`adi`) of person j, and
4. the age at death (`ade`) of person j.

If the person j does not contract the disease, the age at diagnosis `adi` is set to `NA` (missing). In summary, we get the Algorithm 1.

Algorithm 1 Simulation of populations moving in the IDM

for $j=1$ **to** n **do**
  `dob` $\leftarrow t_0^{(j)}$
  calculate event time $T_1^{(j)}$ according to Equation (1)
  simulate type of event that has happened at $T_1^{(j)}$ by Equation (1)
  if event is diagnosis **then**
    `adi` $\leftarrow T_1^{(j)}$
    calculate time $T_2^{(j)}$ of death using Equation (2)
    `ade` $\leftarrow T_1^{(j)}+T_2^{(j)}$
  else
    `adi` $\leftarrow$ `NA`
    `ade` $\leftarrow T_1^{(j)}$
  end if
  write `j, dob, adi, ade` to file
end for

Calculating line integrals

In this section the calculation of the integrals in Equations (1) and (2) is described. The situation in which analytical expressions for these integrals exist, is straightforward. The first simulation in the next section is an example. However, in real world applications, analytical expressions for the integrals are rarely given. For convenience, mathematical functions (e.g., splines) may be fitted to the data and integration is accomplished with the fitted functions. Since the aim of this work is a flexible way of treating the incidence and mortality rates, we assume that the rates are given as numerical values on a regular grid. Here we focus on the most general case, which is characterized by:

1. None of the time-scales t,a, and d is negligible, and
2. the values of i,m_0,m_1 are given as data points only.

For many chronic diseases, we think this is the most relevant case: The mortality rates m_0 and m_1 depend on a and t. Since age is a risk factor for many diseases, the dependency on age is

obvious. Healthier life-style and medical progress in many countries lead to secular trends in m_0 and m_1. In addition, disease duration d is likely to have an impact on m_1 in many chronic diseases. Thus, none of the time-scales is negligible, which is covered in the second and third example.

In the most general case, the integrands i,m_0, and m_1 are given by data points only. We assume that the numerical values are located on a regular grid. The grid is two-dimensional in case of i and m_0, which depend on two time-scales t and a; and the grid is three-dimensional in case of m_1 which depends on t,a, and d.

In event history analysis [2], a useful concept is the *Lexis diagram*, which is a co-ordinate system with axes calendar time t (abscissa) and age a (ordinate). The t-dimension sometimes is referred to as period. Each subject is represented by a line segment from time and age at entry to time and age at exit. Entry and exit may be birth and death, respectively, or entry and exit in a epidemiological study or clinical trial. There are excellent and extensive introductions about the theory of Lexis diagrams (see for example [9,15,16] and references therein), which allows us to be short here. In irreversible diseases, the common two-dimensional Lexis diagram with axes in t- and a-direction may be generalized to a three-dimensional co-ordinate system with disease duration d represented by the applicate (i.e., the z-axis). If a subject does not contract the disease during lifetime, the life line remains in the t-a-plane parallel to the line bisecting abscissa and ordinate. In other words, the life line for the time without disease is parallel to $e_1 := (1,1,0)$ (where the triple (t,a,d) denotes the co-ordinates in time, age and duration direction, respectively). However if at a certain point in time E the disease is diagnosed, the life line changes its direction, henceforth runs parallel to $e_2 := (1,1,1)$.

The situation is illustrated in Figure 2. The life lines of two subjects are shown in the three-dimensional Lexis space. At birth (denoted B_v,$v=1,2$) both subjects are disease-free; both life lines are parallel to e_1. The first subject contracts the disease at E. Henceforth, the life line is parallel to e_2 until death at D_1. The second subject remains disease-free until death at D_2.

Having the concept of the Lexis diagram at hand, we observe that F_1 and F_2 in Equations (1) and (2) are line integrals in the Lexis space. We start with calculating the first failure times T_1. For subject j the associated life line starts at $(t,a)=(t_0^{(j)},0)$. We chose an age $\omega>0$, when it is certain that a transition to one of the states *Disease* or *Death* has occurred, say $\omega=150$ (years). For calculating $F_1^{(j)}$, we trace the hypothetical life line from $B_j := (t_0^{(j)},0)$ to $D_j := (t_0^{(j)}+\omega,\omega)$. Thus, the hypothetical life line has a representation

$$\mathcal{L}_j : B_j+\alpha\cdot(D_j-B_j),\alpha\in[0,1].$$

Following the life line is related to the method of ray-tracing in the field of computer graphics, where efficient algorithms for this purpose exist. In Siddon's algorithm [17], the key idea is to follow $\mathcal{L}_j$ by calculating intersections with volume elements (voxels), which form a regular partition of the Lexis space. Let

$$A_j^\star=\{\alpha^{(j)}(p)|p=1,\ldots,P^{(j)}\}$$

with $0=\alpha^{(j)}(1)<\ldots<\alpha^{(j)}(P^{(j)})=1$ be a parametrization of the points where $\mathcal{L}_j$ intersects the voxel faces plus the start and end points B_j and D_j. Details for the calculation of $A_j^\star$ are described in the supporting information to this article. The parametrization $A_j^\star$ is ideally suited for approximating the integral in Equation (1) by the trapezoidal rule [18]. The reason lies in the fact that in calculating $F_1^{(j)}(\omega)$ the values $F_1^{(j)}\left(t_0^{(j)}+\alpha^{(j)}(p)\omega\right),p=1,\ldots,P^{(j)}$, are a by-product. Algorithm 2 shows the necessary steps.

Algorithm 2 Calculating F_1

1: **for** $j=1$ **to** n **do**
2: calculate $A_j^\star=\{\alpha^{(j)}(p)|p=1,\ldots,P^{(j)}\}$
3: $\ell_1\leftarrow 0$
4: $\tau_1\leftarrow 0$
5: $f_1\leftarrow i\left(t_0^{(j)},0\right)+m_0\left(t_0^{(j)},0\right)$
6: $F_1^{(j)}(\tau_1)\leftarrow 0$

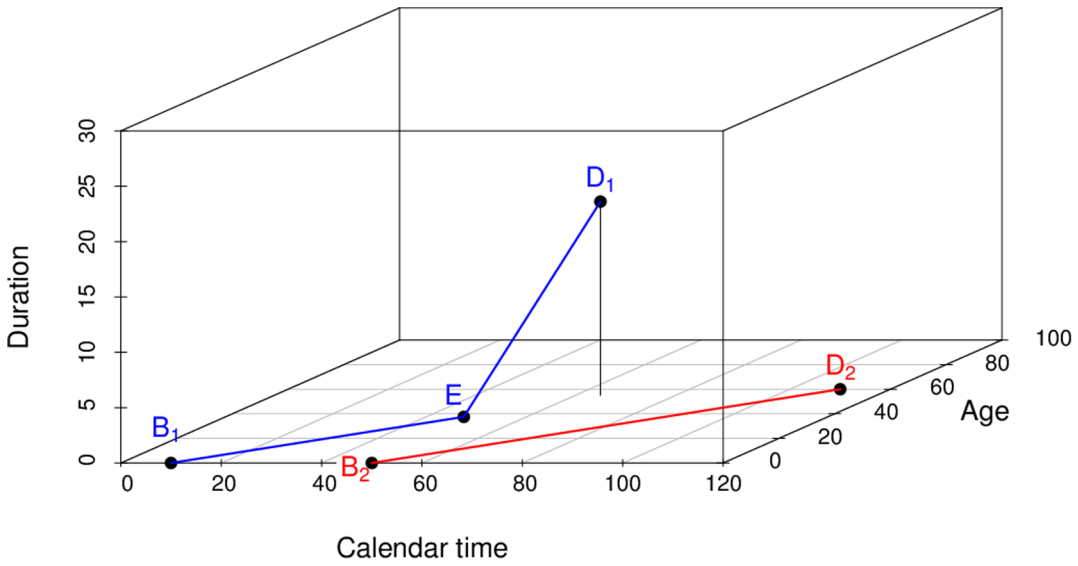

Figure 2. Three-dimensional Lexis diagram with two life lines. Abscissa, ordinate and applicate (z-axis) represent calendar time t, age a and duration d, respectively. The life lines start at birth B_v and end at death D_v,$v=1,2$. The first subject (blue line segments) contracts the disease at E. Then, the life line changes its direction. The second subject (red line segment) does not contract the disease, the life line remains in the t-a-plane.
doi:10.1371/journal.pone.0106043.g002

7: **for** $p=2$ **to** $P^{(j)}$ **do**
8: $\tau_p \leftarrow \alpha^{(j)}(p)\cdot\omega$
9: $f_p \leftarrow i\left(t_0^{(j)}+\tau_p,\tau_p\right)+m_0\left(t_0^{(j)}+\tau_p,\tau_p\right)$
10: $\ell_p \leftarrow \ell_{p-1}+\frac{1}{2}\cdot(\tau_p-\tau_{p-1})\cdot(f_p+f_{p-1})$
11: $F_1^{(j)}\left(t_0^{(j)}+\tau_p\right)\leftarrow 1-\exp(-\ell_p)$
12: **end for**
13: **end for**

Since the values of i and m_0 are given on the voxel grid only, the calculation of $f_p, p=1,\ldots,P^{(j)}$, needs bilinear interpolation of the values of the adjacent voxels [19].

After preparing $F_1^{(j)}, j=1,\ldots,n$, the values of the times $T_1^{(j)}$ can be calculated by the inverse transform sampling method. Since we have $F_1^{(j)}$ calculated at points $\zeta_p := t_0^{(j)}+\tau_p, p=1,\ldots,P^{(j)}$, the inverse transform sampling would yield only those ζ_p. A better accuracy can be obtained by (linear) inverse interpolation [18]. For $t\in(\zeta_{p-1},\zeta_p), p=2,\ldots,P^{(j)}$, let $\xi := \frac{t-\zeta_{p-1}}{\zeta_p-\zeta_{p-1}}$. Then, it holds

$$F_1^{(j)}(t)\approx(1-\xi)\cdot F_1^{(j)}(\zeta_{p-1})+\xi\cdot F_1^{(j)}(\zeta_p).$$

Thus, for $u\in(0,1)$ drawn from a uniform distribution, we find the unique $p\in\{2,\ldots,P^{(j)}\}$, such that $F_1^{(j)}(\zeta_{p-1})\le u<F_1^{(j)}(\zeta_p)$, and set

$$t=\zeta_{p-1}+\left(u-F_1^{(j)}(\zeta_{p-1})\right)\cdot\frac{\zeta_p-\zeta_{p-1}}{F_1^{(j)}(\zeta_p)-F_1^{(j)}(\zeta_{p-1})}$$

as the inverse $\left(F_1^{(j)}\right)^{-1}(u)$.

For those subjects j' who contract the disease, the associated $F_2^{(j')}(\cdot|T_1^{(j')})$ can be derived in a similar way as in Algorithm 2. The associated line segment starts at $(t,a,d)=(t_0^{(j')}+T_1^{(j')},T_1^{(j')},0)$. Again, a hypothetical maximal disease duration ω' is assumed, say $\omega'=80$ (years), such that the line segment ends at $(t,a,d)=(t_0^{(j')}+T_1^{(j')}+\omega',T_1^{(j')}+\omega',\omega')$. Thus, the line segment is parallel to $e_2=(1,1,1)$. The Siddon algorithm computes the corresponding set of intersections with the voxel grid accordingly. The ages $T_2^{(j')}$ at death with disease are obtained from Algorithm 2 mutatis mutandis. The interpolation of m_1 needs to be trilinear.

Examples

This section shows the results of different simulation settings. The associated R [20] code is provided with this article. The first simulation is about a hypothetical chronic disease with rates i, m_0, m_1 only depending on a. The corresponding age-specific prevalence can be calculated analytically, which allows cross-checking the results of the simulation. In the second example, another hypothetical disease is treated with mortality rates depending on t, a, and d. Here we use the ray-tracing approach in the Lexis diagram. Again, the outcomes of the simulation are compared with analytical results. The third simulation is about a relatively rare rheumatic disease. A hypothetical birth cohort of 100000 women is followed from birth to death and examined if the disease is diagnosed. Finally, the last simulation demonstrates applicability in the context of medical decision-making. Two interventions are compared with respect to the outcome *life-years gained*.

Simulation 1: Analytical example

In the first simulation, only one time-scale is involved. Assuming $m_0(a)=m_1(a)=\exp(-10+0.1a)$, and

$$i(a)=\begin{cases}0 & \text{for } a<20\\ \dfrac{0.005}{1-0.005\ (a-20)} & \text{for } a\ge 20,\end{cases}$$

one can show that the age-specific prevalence $p(a)$ is given by $p(a)=0.005(a-20)_+$ [21]. The notation x_+ means the positive part of x: $x_+=\max(0,x)$. The integrals of $i(a), m_0(a)$, and $m_1(a)$ can easily be expressed analytically. Figure 3 shows the age-specific prevalence in a simulated cohort with $n=200000$ subjects. For comparison, the true (analytical) prevalence is depicted as a solid line. The result of the simulation agrees very well with the analytical age profile of the prevalence, which indicates the correctness of the implemented R code.

Simulation 2: Mortality depends on duration

The second simulation is about a hypothetical disease with all time-scales t, a and d playing a role. We aim at calculating epidemiological measures that describe the population of the diseased. For example, age of onset, mean duration of the disease and age at death may be important to plan resource allocation (e.g., inpatient facilities).

In each of sixty consecutive years $t=0,\ldots,59$, 2000 persons are born and followed from birth to death. The incidence of a hypothetical chronic disease is assumed to be $i(t,a)=\frac{(a-30)_+}{3000}$, the age-specific mortality rate of the non-diseased is chosen to be $m_0(t,a)=\exp(-10.7+0.1a+t\ln(0.998))$ and the mortality of the diseased is $m_1(t,a,d)=m_0(t,a)\cdot(0.04(d-5)^2+1)$. In total, 42299 of the 120000 simulated persons contract the disease. The simulated data allows the derivation of important epidemiological measures. For example, the histograms of the age of onset and age at death are shown in Figure 4.

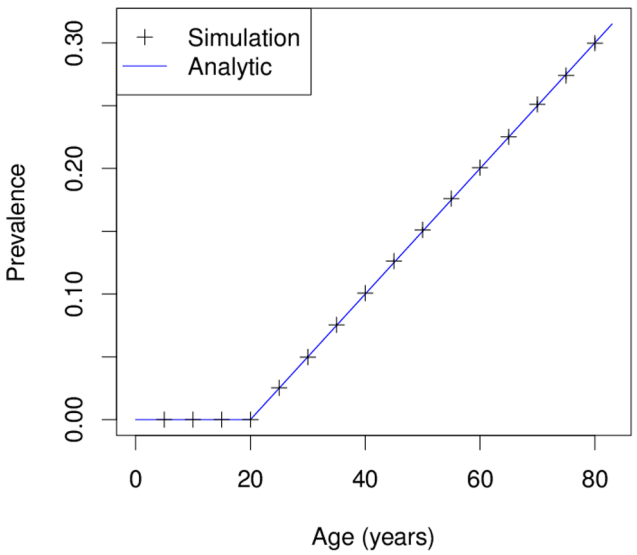

Figure 3. Theoretical and simulated age-specific prevalence. Simulation 1 comprises $n=200000$ persons. The resulting age-specific prevalence (black crosses) is compared to the analytically calculated prevalence (blue solid line). The example shows the very good agreement between the simulation and the theoretical results.
doi:10.1371/journal.pone.0106043.g003

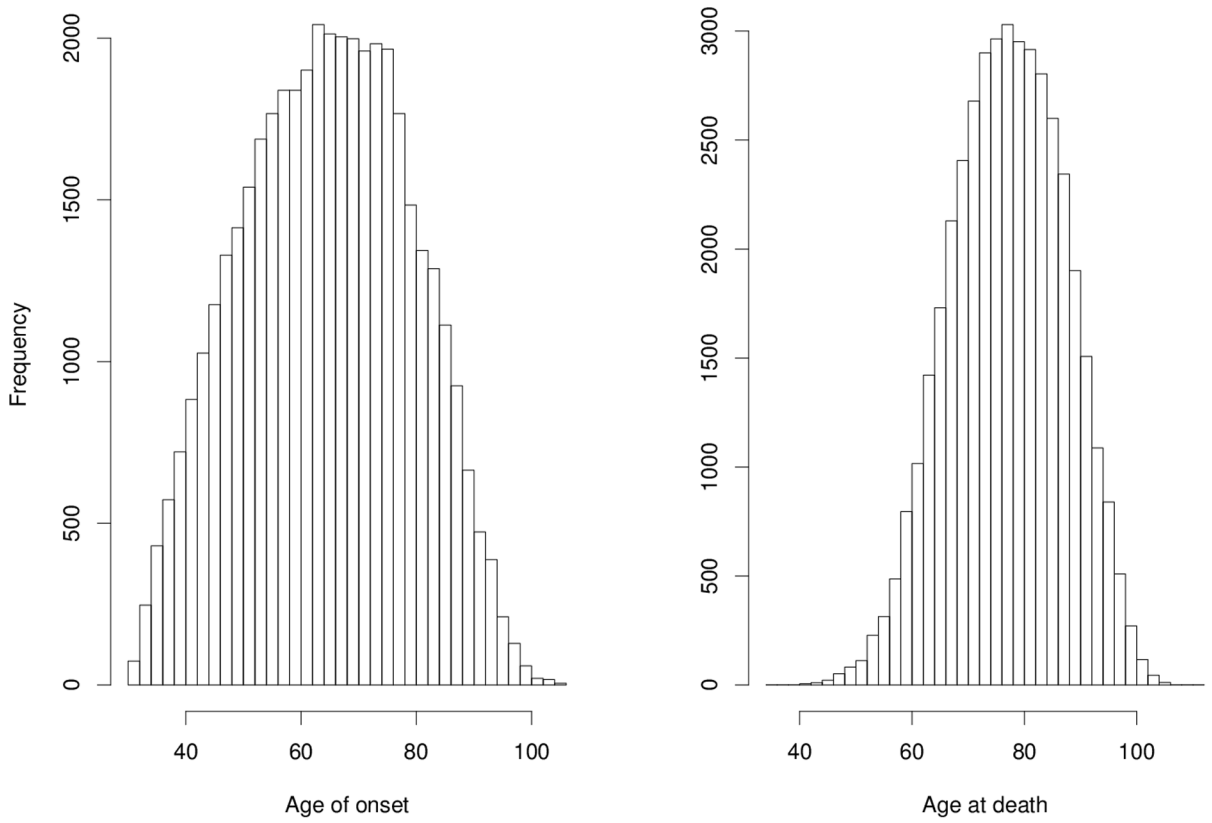

Figure 4. Histograms of the age of onset and age at death in a hypothetical chronic disease. In Simulation 2 the empirical distributions of the age of onset (left) and the age at death of the diseased persons (right) are estimated.
doi:10.1371/journal.pone.0106043.g004

The median age at death of those who contracted the disease is 77.20 (years) whereas the median age at death of those without the disease is 79.67 (years). The median duration of the disease in the 42299 ill subjects is 12.69, the interquartile range is 7.46–17.60 (years).

Finally, we can cross-check the results of the simulation by comparison with an analytical calculation. In year $t = 100$, exactly 76548 persons are alive, 8802 of those having the hypothetical disease. Figure 5 shows the age-specific prevalence at $t = 100$. The black lines indicate the prevalence of several age groups together with 95% confidence bounds as given by the simulation. The blue line represents the prevalence calculated analytically by the exact formula in ([9], Section 7.2). The results agree quite well within the confidence bounds.

Simulation 3: Systemic lupus erythematosus in the UK

The third simulation is about a hypothetical cohort of 100000 women born in the United Kingdom in 1945. The disease under consideration is systemic lupus erythematosus (SLE), which is an autoimmune disorder with significant morbidity and increased mortality. Women are more often affected than men; in terms of age-standardized incidence, the ratio is 5:1 [22]. The peak of the age-specific incidence in British women is located at the age of about 50. Since Somers et al. could not find a significant secular trend in the age-standardized incidence in 1990–1999, we assume that the incidence does not depend on t, [22]. The aim of this example is to study the interplay between age of onset and the mortality of the diseased women in a realistic setting.

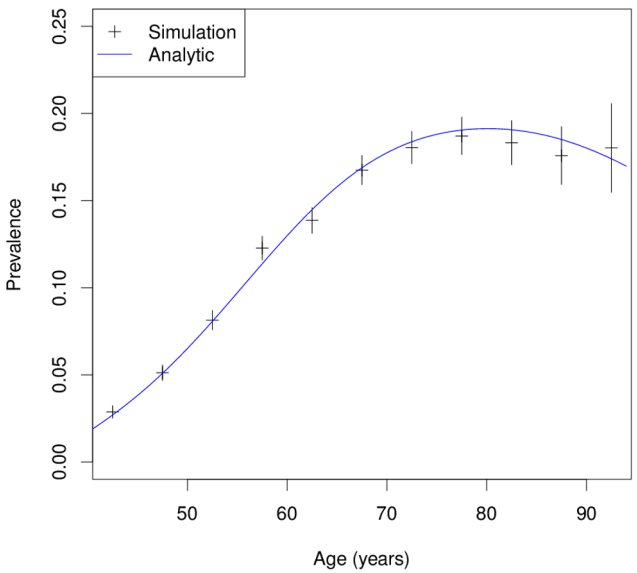

Figure 5. Calculated and simulated prevalence in Simulation 2. If we mimic a cross-sectional study at year $t = 100$, we obtain the age-specific prevalence as indicated by the black bars (with 95% confidence bounds). For comparison the analytically calculated age-specific prevalence is shown as blue line.
doi:10.1371/journal.pone.0106043.g005

Mortality m_0 has been taken from the official life tables of the British Office for National Statistics [23]. The calendar time trend in m_0 has been simplified by assuming that the yearly decrease in mortality is 0.36% for all age groups. This is the geometric mean of the decrease in female mortality taken over all reported age groups in the past 60 years. Mortality m_1 of the women with SLE has been modelled according to [7], which takes into account several covariables: sex, age, duration of SLE, calendar time of diagnosis and geographical region. Unfortunately, an interaction analysis of these factors has not been reported, which forces us to make assumptions. We assume a multiplicative model of the impact of sex, age, region and SLE duration:

$$m_1(t,a,d) = c \cdot m_0(t,a) \cdot H_{\text{age}}(a) \cdot H_{\text{dur}}(d).$$

The constant c reflects the impact of sex and the region, H_{age} and H_{dur} are hazard ratios. The hazard ratio H_{dur} of SLE duration has been interpolated affine-linearly on a logarithmic scale from the published values and backtransformed. The age dependency of $H_{\text{age}}(a)$ has been treated similarly. Due to the weak (and possibly insignificant) effect, the impact of the calendar time of diagnosis has been neglected.

From the 100000 women in the simulation, 513 contract SLE. This corresponds to a lifetime risk of about 0.5%. For comparison, a recent publication about women in the US estimated 0.9% [24]. The study [24] included women with African ancestors, who are known to have a higher risk. The median age of onset in the simulation is 46.07 years, which is in nearly perfect agreement with the result of 46.1 years based on another simulation using the same data but a different method [25]. Table 1 shows the interplay between age of diagnosis and duration of SLE in the simulation.

From Table 1 it is apparent that ten-year survival is not easy to reach. Substantial loss of life time is also indicated by the age at death: Median age at death for the diseased and the non-diseased women is 61.4 and 77.1 (years), respectively. This indicates a considerable loss of lifetime in the population of the diseased compared to the non-diseased. Fortunately, this situation has changed in the last years with better medical care for SLE patients. Especially the introduction of optimized treatment regimes in the last decade lead to an enormous reduction of mortality [26]. These effects have not been included in the simulation.

Simulation 4: Effectiveness of two interventions

We take Simulation 2 as the basecase and compare the effectiveness of two hypothetical intervention programmes A and B with respect to the total gain of life-years. Assume that the primary prevention programme A gradually lowers the incidence rate for all $a \geq 30$ in the calendar years $t \in [0,10]$ by 15% and remains at the 85% level for $t > 10$. Programme B is assumed to reduce the mortality m_1 of the diseased people gradually by 50% starting after six years with the disease. B could be achieved, for example, by listing a new drug on the formulary. After running the simulation, we find that in total the 120000 simulated persons in the basecase, in intervention A, and intervention B, have 9275008, 9348941, and 9351338 life-years, respectively. Hence, intervention A and B yield about 73900 and 76300 life-years more than the basecase. With respect to the chosen outcome parameter, gain of life-years, programme B turns out to be superior to programme A. Thus, primary prevention by programme A, in this example, is less effective than lowering the mortality of the diseased. This result is

Table 1. Age at diagnosis and disease duration of 513 British women with SLE in a simulated cohort.

Age at		SLE duration (years)					
diagnosis		<1	1–2.4	2.5–4	5–9	10–24	≥25
5–34	(n = 134)	12.7%	9.0%	11.9%	14.9%	19.4%	32.1%
35–44	(n = 108)	15.7%	10.2%	14.8%	13.9%	19.4%	25.9%
45–59	(n = 149)	16.1%	14.1%	14.8%	13.4%	26.8%	14.8%
≥60	(n = 122)	29.5%	16.4%	21.3%	23.0%	9.0%	0.8%

Example how to read the table: 108 of the women are diagnosed in the age group 35–44, 13.9% of those die 5–9 years after diagnosis.
doi:10.1371/journal.pone.0106043.t001

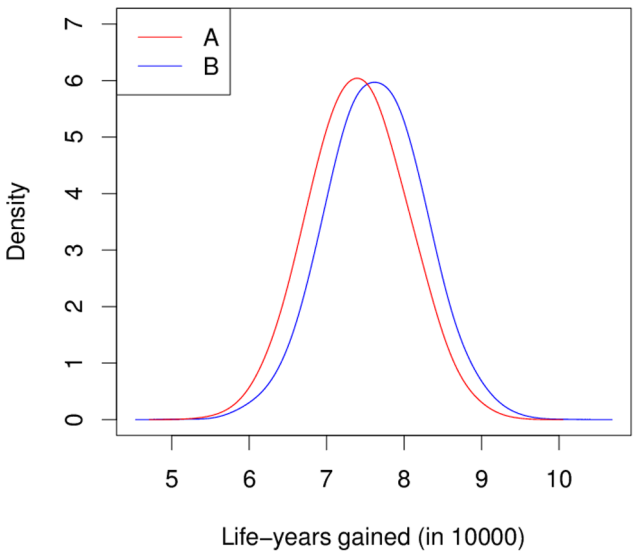

Figure 6. Esimtated densities of the gain in life-years in two intervention programmes. Simulation 4 compares two intervention programmes in a population of 120000 persons. The estimated densities of the total gain in life-years (compared to a basecase) are shown as red and blue curve for programmes A and B, respectively.
doi:10.1371/journal.pone.0106043.g006

hardly predictable without a simulation and demonstrates the usefulness of our algorithm in decision making.

To obtain the empirical distributions of the total gain of life-years in both interventions, 5000 random samples of size 120000 have been drawn (with replacement from the simulated 120000 persons) [27]. Figure 6 shows the resulting densities and the superiority of programme B.

Conclusions

This article is about simulating populations in an illness-death model consisting of the three states *Normal*, *Disease*, and *Death*. The relevant events, diagnosis and death, in the most general case depend on three time-scales: calendar time, age and disease duration.

After birth of a subject in the population, two cases may occur:

1. the subject dies without the disease, or
2. the subject contracts the disease and dies with the disease.

Both situations can be represented in the Lexis space, a common tool in event history analysis. In the first case, the life line is solely located in the time-age-plane. In the second case, the life line changes its direction after onset of the disease, which allows to model the duration of the disease. In many diseases, the duration plays an important role for the mortality. Beside systemic lupus erythematosus as treated above, other diseases such as diabetes [3,4], depression [6] and dementia [5] may serve as examples.

In the most general case, the simulation requires to numerically solve line integrals. Therefore, a synthesis between a ray-tracing technique and numerical integration is exploited. The method provides a fast way to follow the individual life lines in the Lexis diagram. Computation time is an issue, because the number of simulated subjects in the population may be large (several thousands). For example, Simulation 4 takes almost nine minutes (525 seconds) on an Intel i3 personal computer with 3.3 GHz and 8 GB RAM. Simulations 1, 2, and 3 take 40, 170 and 85 seconds, respectively.

Beside the areas mentioned above, we think the method may be applicable in following the fields:

- In epidemiology, the simulation may be used to study the interplay between characteristics in chronic diseases: prevalence, mean disease duration, age of onset, age at death of the non-diseased and diseased population.
- The algorithms may serve as a test bench for estimation methods. For example, in [21] the age-specific incidence is derived from prevalence data. The simulation may be used to study the performance of this and related methods.
- In health economics, the result of our simulation allow the application of cost weights to each subject of the population. For example, in diabetes it is well-known that disease related costs depend on the duration since onset [28]. For each individual the disease related costs may be calculated at a specific point in time. By summing over all subjects, the total costs may be estimated easily.
- Similarly, for many chronic diseases, the health related quality of life depends on the duration of the disease [29]. By assigning utility weights to each individual and summing them up, the total quality-adjusted life-years (QALY) may be calculated.

The last two points are related to health economic modelling. We think that our algorithm may yield a contribution in that domain, because often health economic models are Markov models. Due to memoryless property of Markov models, the dependency of the relevant outcomes on the duration cannot be included directly ([30], Sec. 4.2.3). In the field of diabetes, for example, at least six out of ten important economic models are not capable to accurately account for diabetes duration [31]. This may not be necessary in every research question, but if highly accurate results are needed, modelling the disease duration should be considered.

Supporting Information

File S1 (PDF)

File S2 (ZIP)

Acknowledgments

We dedicate this work to the statistician and demographer Wilhelm Lexis on the 100th anniversary of his death.

Author Contributions

Contributed to the writing of the manuscript: RB. Developed the methods, drafted the text, and made the programming: RB. Critically revised the text, gave important intellectual contributions and final approval of the version to be published: RB SL RFB MS GG.

References

1. WHO (2011). Noncommunicable diseases country profiles. URL http://whqlibdoc.who.int/publications/2011/9789241502283eng.pdf. Last access: May 20st, 2014.
2. Keiding N (2006) Event history analysis and the cross-section. Statistics in Medicine 25: 2343–2364.

3. Carstenson B, Kristensen JK, Ottosen P, Borch-Johnsen K (2008) The Danish National Diabetes Register: Trends in Incidence, Prevalence and Mortality. Diabetologia 51: 2187–2196.
4. Fox C, Sullivan L, Sr RD, Wilson P (2004) The significant effect of diabetes duration on coronary heart disease mortality: the framingham heart study. Diabetes Care 27: 704–708.
5. Rait G, Walters K, Bottomley C, Petersen I, Iliffe S, et al. (2010) Survival of people with clinical diagnosis of dementia in primary care: cohort study. BMJ 341: c3584.
6. Geerlings S, Beekman A, Deeg D, Twisk J, Van Tilburg W (2002) Duration and severity of depression predict mortality in older adults in the community. Psychological Medicine 32: 609–618.
7. Bernatsky S, Boivin JF, Joseph L, Manzi S, Ginzler E, et al. (2006) Mortality in systemic lupus erythematosus. Arthritis and Rheumatism 54: 2550–2557.
8. Beyersmann J, Allignol A, Schumacher M (2011) Competing Risks and Multistate Models with R. Use R! Springer.
9. Keiding N (1991) Age-specific incidence and prevalence: a statistical perspective. Journal of the Royal Statistical Society A 154: 371–412.
10. Kalbfleisch J, Prentice R (2002) The Statistical Analysis of Failure Time Data. John Wiley & Sons, 2nd edition.
11. Aalen O, Borgan O, Gjessing H (2008) Survival and Event History Analysis. Springer-Verlag.
12. Vandenbroucke JP, Pearce N (2012) Incidence Rates in Dynamic Populations. International Journal of Epidemiology 41: 1472–1479.
13. Jewell NP, van der Laan M (2003) Current status data: Review, recent developments and open problems. In: Balakrishnan N, Rao C, editors, Advances in Survival Analysis, Elsevier, volume 23 of *Handbook of Statistics*. pp. 625–642. doi:10.1016/S0169-7161(03)23035-2. URL http://www.sciencedirect.com/science/article/pii/S0169716103230352.
14. Hens N, Aerts M, Faes C, Shkedy Z, Lejeune O, et al. (2010) Seventy-five Years of Estimating the Force of Infection from Current Status Data. Epidemiology of Infections 138: 802–812.
15. Keiding N (1990) Statistical interference in the lexis diagram. Philosophical Transactions of the Royal Society London A 332: 487–509.
16. Carstensen B (2006) Age-Period-Cohort Models for the Lexis Diagram. Statistics in Medicine 26: 3018–3045.
17. Siddon RL (1985) Fast Calculation of the Exact Radiological Path for a Three-Dimensional CT Array. Medical Physics 12: 252–255.
18. Dahlquist G, Björck Å (1974) Numerical Methods. Dover Books on Mathematics Series. Dover Publications.
19. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1988) Numerical Recipes in C. Cambridge University Press.
20. R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.
21. Brinks R, Landwehr S, Icks A, Koch M, Giani G (2013) Deriving age-specific incidence from prevalence with an ordinary differential equation. Statistics in Medicine 32: 2070–2078.
22. Somers EC, Thomas SL, Smeeth L, Schoonen WM, Hall AJ (2007) Incidence of systemic lupus erythematosus in the united kingdom, 1990–1999. Arthritis Care & Research 57: 612–618.
23. ONS (2011). Office for national statistics: Population estimates quinary age groups for uk constituent countries. URL http://www.ons.gov.uk.
24. Crowson CS, Matteson EL, Myasoedova E, Michet CJ, Ernste FC, et al. (2011) The lifetime risk of adult-onset rheumatoid arthritis and other inflammatory autoimmune rheumatic diseases. Arthritis & Rheumatism 63: 633–639.
25. Brinks R, Fischer-Betz R, Landwehr S, Schneider M, Giani G (2013) Stochastic differential equation modeling the prevalence of rare chronic diseases with an application to systemic lupus erythematosus. Statistica Neerlandica 67: 202–210.
26. Chehab G, Fischer-Betz R, Schneider M (2011) Entwicklung von mortalitt und morbiditt beim systemischen lupus erythematodes. Zeitschrift fr Rheumatologie 70: 480–485.
27. Efron B, Tibshirani R (1993) An Introduction to the Bootstrap. Monographs on statistics and applied probabilities. Chapman & Hall/CRC.
28. Caro J, Ward A, O'Brien J (2002) Lifetime costs of complications resulting from type 2 diabetes in the u.s. Diabetes Care 25: 476–481.
29. Sparring V, Nyström L, Wahlström R, Jonsson PM, Östman J, et al. (2013) Diabetes duration and health-related quality of life in individuals with onset of diabetes in the age group 1534 years a swedish population-based study using eq-5d. BMC Public Health 13: 377.
30. Putter H, Fiocco M, Geskus RB (2007) Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 26: 2389–2430.
31. Yi Y, Philips Z, Bergman G, Burslem K (2010) Economic models in type 2 diabetes. Current Medical Research and Opinion 26: 2105–2118.