Going from the incidence rate (or hazard rate) function to the risk function

2018.10.18 for EPIB607

James A. Hanley

Department of Epidemiology, Biostatistics and Occupational Health,
McGill University,
Montreal, Quebec,
H3A 1A2
Canada

tel: +1 (514) 398 6270 fax: +1 (514) 398 4503

email: james.hanley@McGill.CA

Web Page: `http://www.biostat.mcgill.ca/hanley`

Abstract

The risk over a given time span can be calculated as the exponentiated value of the negative of the integral of the incidence density function (or hazard rate function) over that time span. This relationship is widely used but, in the few instances where textbooks have presented it, the derivations of it tend to be purely mathematical. I offer a more intuitive heuristic approach that draws on the conceptualization of a person-year in Edmonds' 1832 definition of the force of mortality, and on the number of replacements in a dynamic population. Similarly I show how the Nelson-Aalen risk estimator can be seen in the context of this historical conceptualization of a person-year, scaled to the experience of a dynamic population of (constant) size 1.

Note: JH submitted this to Epidemiology in 2012 and 2013, to be told that it should be written as a 'K-M vs. N-A (which is better) contest'.

He submitted it to Lifetime Data Analysis in 2016, to be told that the use of the integral as the mean of a Poisson random variable, along with the concept of a dynamic population, was well know in the renewal-process and counting-process literatures, and that it was written more for epidemiologists than statisticians.

If readers find the current version useful, or have suggestions for how to improve it,

please contact JH.

1 Introduction

In the abstract, cumulative incidence is – in statistical terminology – a *parameter*: the proportion of a theoretical cohort – all members of which are candidates initially – that, over a specified period or span of time [or age], leaves the initial state and enters the state of interest. Typical applications are the 30-day mortality rate, and the 'x'-year risks of various illnesses.

In some instances, its empirical counterpart, which statisticians refer to as an estimate, can be calculated directly as an proportion; in others, it can only be arrived at indirectly, for example as the complement of the product of conditional 'survival' probabilities in an actuarial or medical life-table framework[1]. These indirect computations are described in all epidemiology texts and the – quite intuitive – steps are easily recalled. Less often presented, and less well mastered, is the method based on the incidence density (ID) function or hazard function over that specified time span. This is understandable, as the formula, involving the exponential function and an integral of the ID function, is less transparent. As an simple example, suppose the ID of needle-stick injuries was assumed to be constant (at, say, 0.0975 per intern-month[2]) over the 12-month time-span of interest. Without resorting to one of these textbooks, few end-users could quickly – and confidently – recall and use the 'exponential' formula to convert this incidence function to an incidence proportion or risk.

Yet, this link between the ID function and risk is increasingly used. The

Nelson-Aalen estimator, now offered alongside the familiar Kaplan-Meier estimator in software packages, is based directly on it. In both non-parametric and parametric statistical approaches, the link is used to calculate profile-specific x-year risks,[3] risk differences, and numbers needed to treat.[4] And, it is used in the reverse direction to assess (non)proportionality of hazards via log[survival] plots.

The primary objective of this article is to demystify the 'exponential' formula, by taking advantage of Edmonds' conceptualization of a person year as "one person 'constantly living' for one year" (today termed a dynamic population of constant size 1[5,6]). I clarify the meaning of the integral, and exploit it, along with of a little-used property of the Poisson distribution, to arrive at the formula. A secondary objective is to provide a different heuristic for the Nelson-Aalen estimator of a survival probability or risk. Some of these concepts are implicit in the 'counting process' approach to survival analysis; we attempt to make them explicit and accessible, and to use them didactically. A further objective is to connect some dots, and some terminology, in the disciplines of demography, epidemiology and survival analysis.

2 Definitions, distinctions, and links

We have recently traced the basic concepts and terms, [JRSS]. As far as we can determine, the term 'force' of mortality was introduced to demography by the actuary Edmonds in 1832; the term hazard was introduced to statistics

by Davis in 1952, and its first formal appearance is in the 1959 Technometrics paper by Zelen. The *term* 'incidence density' was introduced to epidemiology by Miettinen in 1976[7] to describe the quantity with dimension $time^{-1}$, and to distinguish it from the 'proportion-type' incidence rate or risk.

Not surprisingly, the distinction between the *concepts* of 'rate' (in the ID, or force of mortality sense) and 'risk' is a very old and well described one[8,9]. But few of the derivations of the link between the 'hazard rate' and 'risk' give any historical or cross-disciplinary references.

2.1 Incidence density (ID), and ID functions

Incidence density (ID) refers to the rate of transition from an initial (usually, but not necessarily, health) state to a different state of interest:

> Incidence density ("force of morbidity" or "force of mortality") – perhaps the most fundamental measure of the occurrence of illness – is the number of new cases divided by the population-time (person-years of observation) in which they occur.

Each (age-specific) value in the ID function in Figure 1 was computed from the *dynamic population* experience in which "a population (...) with turnover of membership moves over (7 years of) *calendar time*, with all members being candidates throughout (so that the transition at issue is among the mechanisms of removal of individuals from the candidate population)." Alternatively, an ID can can be derived from a *cohort* experience, "in which

Table 1: Incidence density (ID), calculated for (successively smaller) intervals, of width Δt, centered on 3 different ages

	t = age 39.25*			t = age 59.25			t = age 79.25		
Δt	PT*	Deaths	ID	PT	Deaths	ID	PT	Deaths	ID
1 year	31.255	55,590	177.9	19.520	177,133	907.5	9.578	486,785	5,082.3
1 month	2.605	4,629	177.7	1.626	14,772	908.6	0.798	40,567	5,082.5
1 week	0.601	1,068	177.7	0.375	3,409	908.6	0.184	9,362	5,082.5
1 day	0.086	152	177.7	0.053	486	908.6	0.026	1,334	5,082.5

*PT: Population-time (1 unit = 1 million person-years) ID Units: deaths / 100,000 person-years

Based on population-sizes and numbers of deaths, USA 2000-2006.
Source: Human Mortality Database, `http://www.mortality.org/`
*Technically, persons are only *exactly* 39.25 for a moment (infinitesimal, since a moment has no duration), so we can only consider the calculation over an interval $(t - \Delta t/2,\ t + \Delta t/2)$, of width Δt, that includes $t = 39.25$.

an enumerable set of individuals, all candidates initially, moves over the risk period.' As in the dynamic case, the cohort experience can be segregated by age or time to yield an ID function, rather than a single ID value.

The concept of a 'hazard' is typically introduced in the context of a cohort. Arguably, the 'mathematical limit' used to define it is more easily understood via the ID in a dynamic population experience. Table 1 gives three examples of the 'ID in the limit' at (deliberately selected to be a non-integer) ages. An ID function that yields a continuous curve shortens otherwise tedious calculations of annuities and of x-year cumulative incidence rates (risks).

2.2 Cumulative incidence rate, or risk

A cumulative incidence proportion (see Introduction) used as the *probability* of transition for an *individual* is usually referred to as a *risk*. William

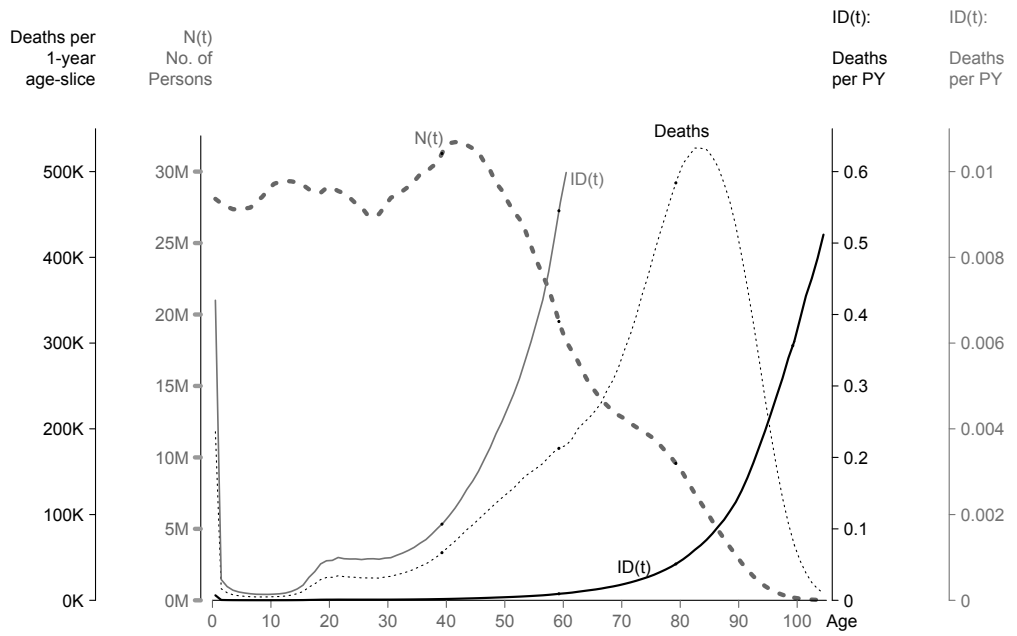

Figure 1: Age structure of, and age-specific numbers of deaths recorded in, the (dynamic) USA population observed over the period January 1, 2000 to December 31, 2006, along with the age-specific death rates (incidence densities) derived from them. Source: Human Mortality Database, `http://www.mortality.org/`. For each (continuous) value of age (t), shown as the thicker dotted line – with vertical axis in millions (M) – is $N(t)$, the number of persons who were 'exactly t years of age' on *some* date in the 7 calendar years. Thus the numbers of person-years lived in any age-interval is the integral of (area under) the $N(t)$ curve over the interval in question. Most residents contributed 7 person-years each to the overall total of 2,000 million person-years; some contributed fewer – mostly to either the younger or older end of the person-years distribution. The numbers of deaths for any age-interval is the integral of the 'deaths per 1-year-of-age time slice' curve (thinner dotted line, axis on 100s of thousands) over the interval in question. The solid black curve (vertical axis in increments of 0.1) is the full, and the solid grey curve the 'below age 60' portion (vertical axis in increments of 0.002) of the $ID(t)$, or force of mortality or hazard rate function. The $ID(t)$ function ranges from a nadir of 0.000014 year^{-1} at approx. $t = 10$, to 0.51 year^{-1} at age $t = 105$.

Farr[8, p2] made this same distinction[10, p229] between an expected proportion in an aggregate, and the probability for an individual. He distinguishes patients 'in two lights', in '*collective masses, when general results can be predicted with certainty*' or '*separately, when the question becomes one of probability.*'

2.3 The formula linking ID and risk

According to Chiang[11,p198], the equation that converts a smooth $ID(t)$ function into a risk "has been known to students of the lifetable for more than two hundred years. Unfortunately, it has not received much attention from investigators in statistics, although various forms of this equation have appeared in diverse areas of research".

Epidemiologic coverage of it begins in 1976.[5] Miettienen's worked example addressed the 30 year risk of bladder cancer for a 50 year old man, assuming that "without bladder cancer he would survive that period." We address the 20 year risks of death from *any* cause for 39.25, 59.25 and 79.25 year olds, making competing risks irrelevant. Thus, the formula can be used directly: the cumulative incidence-rate (CIR) for the age span a' to a'' is

$$CIR_{a',a''} = 1 - \exp\left[- \int_{a'}^{a''} ID(a)da \right]$$

2.4 Subsequent coverage of the link

Miettinen[5,13] merely cites Chiang[12] for the source of the equation; Morgenstern et al.[14] tell us that 'a little calculus' leads to it. Rothman's 1986 textbook[15, *pp*29−31] is the only epidemiological textbook I am aware of that formally derives it. The derivation – with $S(t)$ as the solution of a differential equation – is the same one typically used in survival analysis textbooks, and in the 1980 article[14] This was also the approach used by Edmonds([16, *p xvii*]) in 1832 and, (implicitly) by Lambert in 1765 and Bernoulli in 1776[17].

The popular textbook[18,19] gives the discrete (i.e., summation) version, stating that it is sometimes referred to as the *exponential formula*. It illustrates it using a numerical example, in which the Kaplan-Meier estimator yields a 19-year risk of 0.56. while the exponential estimator yields 0.52. The reader is left wondering which is an approximation to which.

The introductory textbook[20, *pp* 47−51] uses heuristic arguments, but does not show the full-blown formula. Instead it presents "the simplest formula to convert an incidence rate to a risk" : Risk $\approx$ Incidence rate $\times$ Time, which gives a very close approximation in the two worked examples presented. The warning that the product "can exceed 1" in some instances raises, but does not answer, the question of what the product truly means.

Rather than advocate an approach in which the product is sometimes 'close to the numerical value of risk' and sometimes not, I explain what the product means – it has the same meaning whether it is large or small – and that a simple transformation of it can always be interpreted as a risk. In the

next section, I give the product (or more generally, the sum of products, i.e. the integral) in this 'exponential formula' a concrete meaning.

3 A heuristic inspired by Edmonds

To do so, I take up Edmonds' concept of *a given number of persons constantly living*,[16] which he used in originating the term 'force of mortality.' I begin with a simpler shorter-term example, where ID is constant over the entire span in question, before addressing the general time-varying ID case.

3.1 Constant-over-entire-timespan ID

How to convert an incidence density of 0.0975 [1] (first) percutaneous injuries per intern-month[1] —assumed *constant* over a span of 12 months – into a 12-month cumulative incidence (proportion-type) rate or risk?

As Edmonds did, take the 'given number of interns' to be one (1). Imagine a 'chain', starting at $t' = 0$ and extending for 12 months until $t'' = 12$. The chain is begun with a randomly selected never-injured intern, who continues until he/she either reaches 12 months or is injured before then. If the latter, and it occurs at age t, he/she is immediately replaced by another a randomly selected never-injured intern. The chain proceeds, 'with further replacements as needed,' until $t'' = 12$. From t' to t'', the 1 constantly-serving candidate constitutes a dynamic population with a constant membership of 1.

The number of replacements required is a random variable, with possible

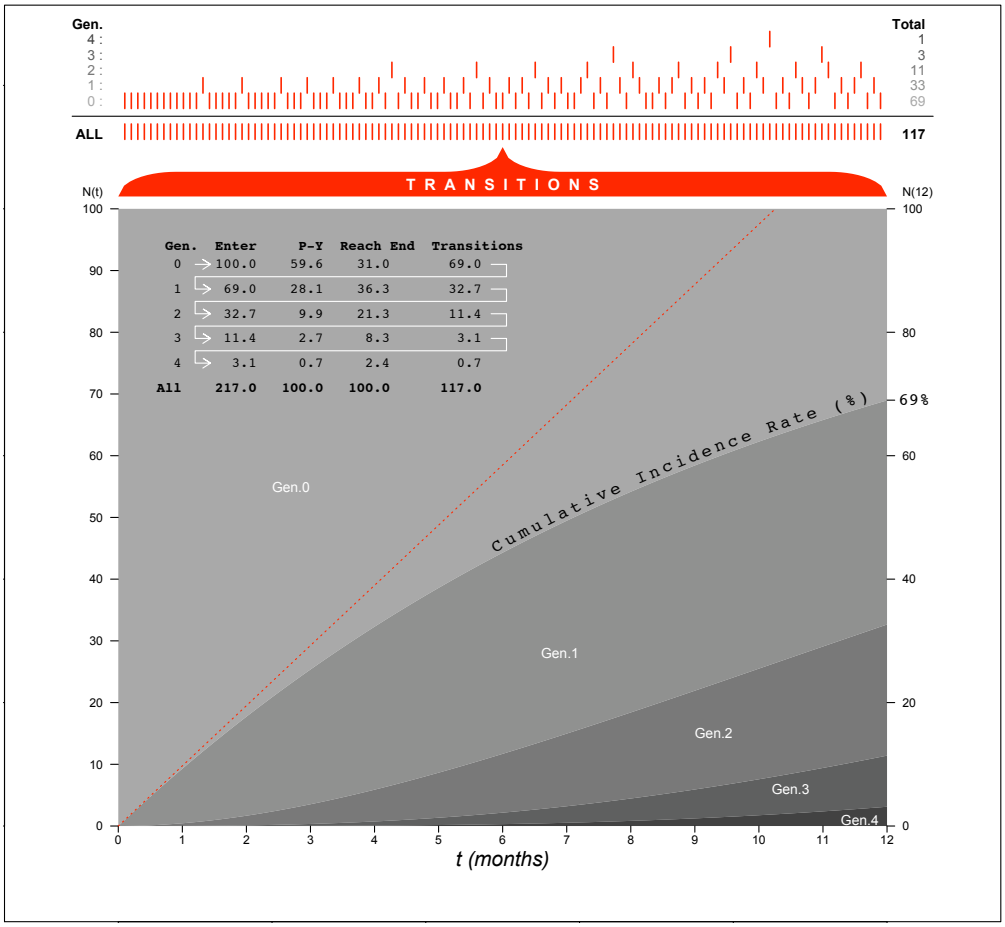

Figure 2: An average of 1.17 transitions (percutaneous injuries) in 1 intern-year (I-Y) of experience (117 in 100 I-Y), so that ID = 1.17 $year^{-1}$. 100 'chains' start at $t = 0$ (the 100 chains are represented by 100 horizontal lines, so close to each other that the total person time appears as a rectangle 100 interns high by 12 months wide); each chain continues for 12 months, each using as many replacements (Gen. 1, 2, ...) as necessary to complete the chain. The different shaded areas represent the population-time for generations 0, 1, ... The proportion of chains that are completed using the initial (Gen. 0) intern is $\exp[-1.17] = 0.31$, i.e., 31%, so the 1-year risk is 100% - 31% = 69%. The proportion of chains in which, by time t, the initial (Gen. 0) intern has been replaced, i.e., the cumulative incidence rate up to time t, is $1 - \exp[-ID \times t] = 1 - \exp[-(\text{integral up to time } t)]$ The straight line (the product of ID and time, scaled up by 100) involves a constant number of candidates at each time point, and thus overestimates the cumulative incidence rate – substantially so as generation 0 is replaced. The numbers of transitions do not sum exactly to 117 because of rounding.

values 0, 1, 2, ... Its expected value (mean) is $\mu = 0.0975\ m^{-1} \times 12\ m = 0.00039\ h^{-1} \times 3000\ h = 1.17$ first injuries. Readers will recognize μ as integral of the $ID(t)$ function over the 12-month age-span. The probability that the chain is completed by the same intern who initiated it is the probability that 0 replacements are required. The probability that it is not is the complement of this 'survival' probability. Since the number of replacements (transitions, first injuries) in the 12 months is a Poisson random variable, the probability that the chain *is* completed by the same intern who initiated it is the Poisson probability of observing 0 events when 1.17 are expected, i.e., as $\exp[-1.17] = \exp\left[-\int_{t'}^{t''} ID(t)dt\right] = 0.31$. The probability that this intern fails to complete the chain, i.e., *is injured before the 12 month period ends* is $1 - \exp\left[-\int_{t'}^{t''} ID(t)dt\right] = 1 - 0.31 = 0.69$. Thus the *12-month risk* of injury is 69%.

Fig 2, modeled after Fig 1 in Miettinen,[5] gives the expected values for a total of 100 separate such chains, and shows why *the product of ID and time (the 1.17, the integral) is not a risk, but rather an expected number of events (transitions, turnovers, injuries) in a dynamic population of size 1.* To provide 100 intern-years of service, an average of 217 interns is required. Of the 100 who began the chains (the average service of these 100 in 'generation 0' is 0.596 P-Y per intern) 31 complete them and 69 do not. Thus, the 12-month risk is 69%. On average, of their 69 replacements (generation 1), 36 complete the chains and 33 do not; and so on, so that in all – over the initial and replacement generations, totaling 100 P-Y – 117 do not and 100 do.

The proportion of chains in which, by time t, the 'Gen. 0' intern has

been replaced, is $1 - \exp[-ID \times t] = 1 - \exp[-(\text{integral up to time } t)]$. The straight line (the product of ID and time, scaled up by 100) involves a constant number of candidates at each time point, and thus overestimates the cumulative incidence rate – substantially so as 'gen. 0' is replaced.

Table 4.2 and Fig. 4.3 of Rothman[20] show a 20-year incidence proportion, but using an ID of 0.011 yr^{-1}, so that the expected number of transitions in a dynamic population of 1 is $0.011\text{yr}^{-1} \times 1 \text{ yr} = 0.22$. Rothman's curve is identical to the first $0.22/0.0975 = 2.3$ months of the curve for the percutaneous injuries.

The expected numbers of 'cumulative deaths' column in Rothman's Table 4.2 can be (and probably were) arrived at using the 'exponential' formula

$$1000 \times \{\ 1 - \exp[-\ 0.011\text{yr}^{-1} \times\ (\text{number of years})]\ \}.$$

The $0.011\ \text{yr}^{-1} \times$ (number of years) is the integral of the ID function, i.e., the expected number of transitions, over the number of years in question.

3.2 Varying-over-the-timespan ID

In chapter 3 in the first (2002) edition of his introductory textbook, Rothman worked though the calculation of 85-year risk of dying of a motor-vehicle injury – assuming (for didactic purposes) one had not died of some other cause in the meantime.

Consider now with the 20-year risk of death from any cause for a person

aged $a' = 79.25$, based on the – clearly non-constant – ID function shown in Figure 3(A).

In his first (2002) edition, Rothman consider the expected number of motor-vehicle injury deaths in a continuous 1-person chain (dynamic population) is

Again, as Edmonds did, we imagine a 1-person 'chain' that starts with a randomly selected living person aged $a' = 79.25$ and extends – with 'with further replacements as needed' – for 20 years until $a'' = 99.25$.

The number of replacements (deaths) in the 1-day-wide interval centered on age t is a Poisson random variable with expected value $ID(t) \times 1[person] \times (1/365.25)[year]$. The sum of 7305 independently distributed daily Poisson random variables, each with a different expected value, is again a Poisson random variable with expected value equal to the sum of these daily expected values. This sum – effectively the integral, from 79.25 to 99.25, of the ID function in Figure 1(B) – is $\mu = 3.22$ transitions/replacements/deaths. The sum of a number of Poisson random variates is again a Poisson variate. Thus, we can calculate the probability that the chain *is* completed by the same person who initiated it, as the Poisson probability of observing 0 events when 3.22 are expected, i.e., as $\exp[-3.22] = \exp\left[-\int_{79.25}^{99.25} ID(t)dt\right] = 0.04$. The 20-year risk is its complement, namely 1 - 0.04 = 0.96, or 96%. The righmost panel of Fig 3(B) shows risk curve for spans shorter than 20 years.

To obtain the 20-year risk for a person aged 59.25, we calculate 1 minus the Poisson probability of observing 0 events when 0.47 events are expected,

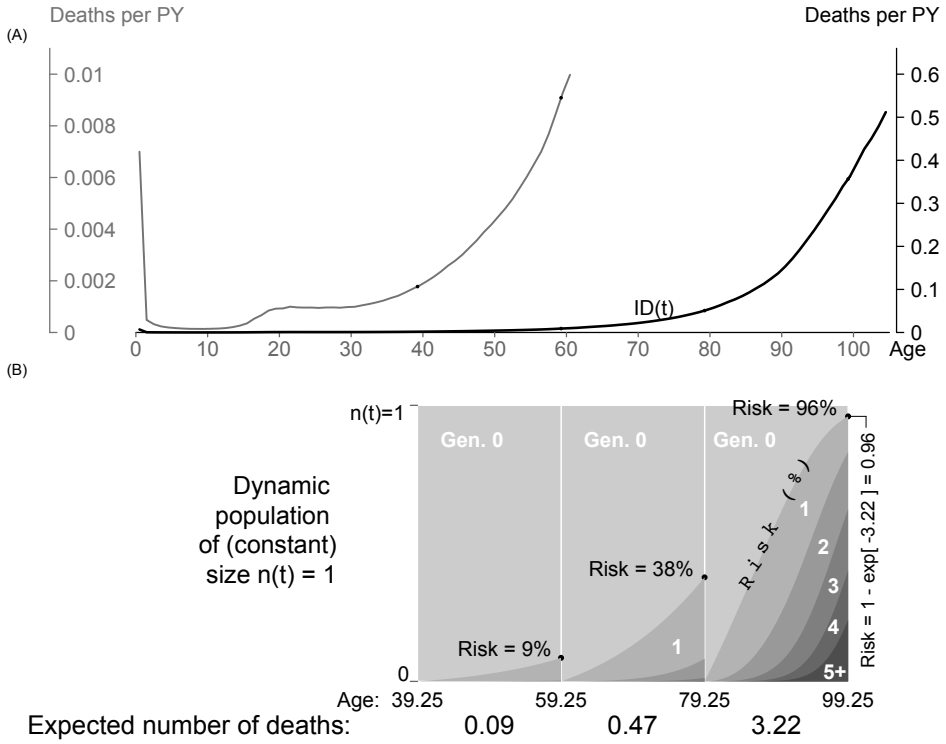

Figure 3: **(A)** The death rate (incidence density, or force of mortality) as a function of age (details in Figure 1) – lighter curve (vertical axis on left) is the 'below age 60' portion – (B) 20 year risks calculated from the ID function. Over the span from 79.25 to 99.25 years, the integral of the ID function, which can be successively approximated by a sum of products, each of the form $ID(t_{mid}) \times 1(\text{person}) \times \Delta t$ (with each ID evaluated at the midpoint of a very small time interval of width Δt) is 3.22. Note that the '1' person in the '$1(\text{person}) \times \Delta t$' amount of population-time experience does not appear explicitly in the usual formulation - the integral is usually written as $\int ID(t)\Delta t$ rather than $\int ID(t) \times \{1 \times \Delta t\}$.

The expected numbers of deaths "if 1 person (not necessarily the same person for the entire span) were constantly living for a 20-year span" (i.e., in a dynamic population of size 1) are shown for 3 selected spans. As in Figure 2, the different shaded areas represent the population-time for generations 0, 1, ... The 20-year risk for a person 79.25 years old is the (Poisson) probability that at least one replacement is observed, if 3.22 replacements are expected. Over the combined 60-year span from 39.35 to 99.25 years of age, a total of 0.09+0.47+3.22 = 3.78 replacements are expected, so the 60-year risk is $1 - \exp[-3.78] = 97.7\%$.

i.e., $1 - \exp[-0.47] = 0.37$. The 40-year risk for a person aged 59.25 is 1 minus the Poisson probability of observing 0 events when 3.22+0.47 = 3.59 events are expected, i.e., $1 - \exp[-3.69] = 0.98$, or 98%.

Again, as Edmonds did, we imagine a 1-person 'chain' that starts with a randomly selected living person aged $a' = 79.25$ and extends – with 'with further replacements as needed' – for 20 years until $a'' = 99.25$.

The number of replacements (deaths) in the 1-day-wide interval centered on age t is a Poisson random variable with expected value $ID(t) \times 1[person] \times (1/365.25)[year]$. The sum of 7305 independently distributed daily Poisson random variables, each with a different expected value, is again a Poisson random variable with expected value equal to the sum of these daily expected values. This sum – effectively the integral, from 79.25 to 99.25, of the ID function in Figure 1(B) – is $\mu = 3.22$ transitions/replacements/deaths. The sum of a number of Poisson random variates is again a Poisson variate. Thus, we can calculate the probability that the chain *is* completed by the same person who initiated it, as the Poisson probability of observing 0 events when 3.22 are expected, i.e., as $\exp[-3.22] = \exp\left[-\int_{79.25}^{99.25} ID(t)dt\right] = 0.04$. The 20-year risk is its complement, namely 1 - 0.04 = 0.96, or 96%. The righmost panel of Fig 3(B) shows risk curve for spans shorter than 20 years.

To obtain the 20-year risk for a person aged 59.25, we calculate 1 minus the Poisson probability of observing 0 events when 0.47 events are expected, i.e., $1 - \exp[-0.47] = 0.37$. The 40-year risk for a person aged 59.25 is 1 minus the Poisson probability of observing 0 events when 3.22+0.47 = 3.59

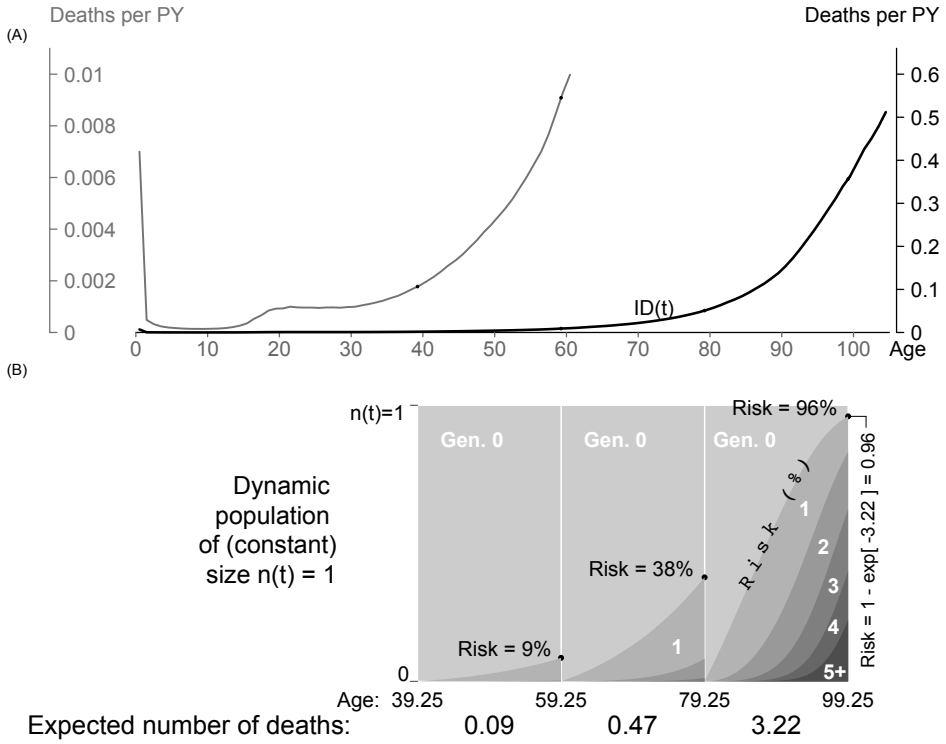

Figure 4: **(A)** The death rate (incidence density, or force of mortality) as a function of age (details in Figure 1) – lighter curve (vertical axis on left) is the 'below age 60' portion – (B) 20 year risks calculated from the ID function. Over the span from 79.25 to 99.25 years, the integral of the ID function, which can be successively approximated by a sum of products, each of the form $ID(t_{mid}) \times 1(\text{person}) \times \Delta t$ (with each ID evaluated at the midpoint of a very small time interval of width Δt) is 3.22. Note that the '1' person in the '$1(\text{person}) \times \Delta t$' amount of population-time experience does not appear explicitly in the usual formulation - the integral is usually written as $\int ID(t)\Delta t$ rather than $\int ID(t) \times \{1 \times \Delta t\}$.

The expected numbers of deaths "if 1 person (not necessarily the same person for the entire span) were constantly living for a 20-year span" (i.e., in a dynamic population of size 1) are shown for 3 selected spans. As in Figure 2, the different shaded areas represent the population-time for generations 0, 1, ... The 20-year risk for a person 79.25 years old is the (Poisson) probability that at least one replacement is observed, if 3.22 replacements are expected. Over the combined 60-year span from 39.35 to 99.25 years of age, a total of 0.09+0.47+3.22 = 3.78 replacements are expected, so the 60-year risk is $1 - \exp[-3.78] = 97.7\%$.

events are expected, i.e., $1 - \exp[-3.69] = 0.98$, or 98%.

In his first (2002) edition, Rothman calculated the expected number of motor-vehicle injury deaths in a continuous 1-person chain (dynamic population) is

$$\frac{4.7}{105\text{Y}} \times 15\text{Y} + \frac{35.9}{105\text{Y}} \times 10\text{Y} + \frac{20.}{105\text{Y}} \times 20\text{Y} + \frac{18.4}{105\text{Y}} \times 20\text{Y} + \frac{21.7}{105\text{Y}} \times 20\text{Y} = 0.016335.$$

and so we arrive at the 85-year risk of $1 - \exp[-0.016335] = 0.016$ or 1.6% with fewer calculation steps that using the method he employed.

4 Approximation to Risk

From the expected value of 0.09 in Figure 3(B), the 20-year (all-cause mortality) risk for a person aged 39.25 is $1 - \exp[-0.09] = 0.086$ or 8.6%. This example, and the one involving the expected value of 0.016335, reflect the fact that, if the expected value (E), is small, so that $1 - \exp[-E] \approx E$, then

$$\text{Risk}_{a',a''} \approx \text{Expected no. (E) of events in } (a', a'') \text{ span, if E is small.}$$

The following shows that the percentage over-estimation in using the approximation $\text{Risk}_{approx} = E$ increases with increasing E, and is close to $(50 \times E)$%.

Expected no. of events, E:	0.02	0.05	0.10	0.20	0.30	0.50	1.00
Risk $= (1 - \exp[-E])$:	0.0198	0.049	0.095	0.181	0.259	0.393	0.632
% by which E overestimates Risk:	1	3	5	10	16	27	58

A large E can arise from a low rate over a longer interval, (e.g., 0.47 from mortality rates in the age span 59.25 to 79.25) or higher ones over a shorter one (e.g. 0.37 from mortality rates in the age span 99.25 to 100.25). Farr was 'aware that when the number of deaths is small, relative to the population studied, both measures approach each other numerically.'[9]

5 The Nelson-Aalen estimator

Although still not covered in most epidemiology texts, the Nelson-Aalen estimator of the survival function[21] is included in most software packages along with the Kaplan-Meier one, and is increasingly found in the medical literature. Both estimators are calculated for survival data that have been reduced to J *very narrow* event-containing sub-intervals of the full $[0, t]$ interval of interest. Interval j is defined by distinct event-time t_j. Intervals in $[0, t]$ that don't contain events are 'non-contributory' and thus ignored. The j^{th} riskset is the set the 'candidates' (n_j in all) just before the event(s) in interval j. Some s_j 'survive' event-containing interval j, while the remaining d_j do not.

In the Kaplan-Meier Product Limit estimator, each of the J empirical conditional probabilities $s_1/n_1, \ \ldots, \ s_J/n_J$ is treated as a surviving fraction of the previous fraction, and so, ultimately, the estimator is simply the overall product of these:

$$\widehat{S(t)}_{KM} = \frac{s_1}{n_1} \times \cdots \times \frac{s_J}{n_J} = \prod_j \frac{s_j}{n_j} = \prod_j \left\{1 - \frac{d_j}{n_j}\right\}.$$

The Nelson-Aalen one is often merely presented, without justification, as

$$\widehat{S(t)}_{NA} = \exp\Big\{ -\sum_j \frac{d_j}{n_j} \Big\}.$$

Sometimes it is accompanied by the statement that "the Kaplan-Meier Product Limit estimator is an *approximation to*" to the Nelson-Aalen one. The following algebra shows that the *reverse* also holds true: when *each* d_j/n_j is small, then $1 - d_j/n_j \approx \exp[-d_j/n_j]$, and so

$$\widehat{S(t)}_{KM} = \prod_j \Big\{ 1 - \frac{d_j}{n_j} \Big\} \approx \prod_j \Big\{ \exp\Big[-\frac{d_j}{n_j} \Big] \Big\} = \exp\Big\{ -\sum_j \frac{d_j}{n_j} \Big\} = \widehat{S(t)}_{NA}$$

But the Nelson-Aalen estimator of the survival function can also be justified in its own right, as the empirical version of the exact link between ID and risk. This in turn can be though of as the Poisson probability ($\exp[-\mu]$) of 0 events, where μ is the number of events that would be expected if the empirical ID function ($\widehat{ID}$) were applied to a dynamic population with a constant membership of one ("one person constantly living"), over the timespan $(0, t)$. The Nelson-Aalen formula is unveiled by examining the structure of the integral, $\mu = \int_{u=0}^{u=t} \widehat{ID(u)} du$. The integrand takes on J positive values $\widehat{ID}_1$ to $\widehat{ID}_J$ inside the J small event-containing intervals, and the value $\widehat{ID(t)} = 0$ everywhere outside of these intervals. If the width of interval j is Δt, then for all values of u within interval j, the fitted ID is $\widehat{ID(u)} = \frac{d_j}{n_j \times \Delta t}$.

Thus, the overall integral is a sum of J non-zero integrals:

$$\mu = \sum_j \left\{ \int \widehat{ID_j(u)} du \right\} = \sum_j \left\{ \frac{d_j}{n_j \times \Delta t} \times \Delta t \right\} = \sum_j \left\{ \frac{d_j}{n_j} \right\}.$$

Fig 4 illustrates the heuristics using data on the frequency of IUD discontinuation because of bleeding.[21, p5] The fitted number of transitions (discontinuations), $\hat{\mu} = \sum_1^9 (d_j/n_j) = 1.25$, is the number of transitions we would expect in a dynamic population of size 1 followed for 107 weeks. This fitted number is obtained by *scaling* the observed population-time so that there is always 1 candidate, and scaling the numbers of transitions accordingly. The 107-week risk of discontinuation is therefore $1 - \exp[-1.25] = 71\%$.

Terminology: Confusingly, different software packages report by different entities as 'Nelson-Aalen' estimates. Sometimes the term is used to describe $\widehat{\mu}$, the *number of events* for a 1-person dynamic population, and sometimes the *risk* $1 - \exp[-\widehat{\mu}]$. As we saw above, the two are sometimes numerically close, sometimes not. Referring to $\widehat{\mu}$, i.e., the sum of products or integral, as the 'integrated hazard' or the 'cumulative hazard', avoids confusion. But – as Rothman et. al [18,19] lament – the term "cumulative incidence" can foster confusion: does the term refer to the integral or the risk? To avoid this possibility, I have used Miettinen's term "cumulative incidence *rate*," but also tried to ensure that readers know when I use the word "rate" in the 'proportion' sense.

6 Closing remarks

I have described four aspects of the expected number of replacements in Edmonds' '1 person constantly living' or Miettinen's 'dynamic population of (constant) size 1'. (1) It provides a heuristic for the integral in a central epidemiologic formula. (2) It has a different conceptual meaning than 'risk' – even if it sometimes provides a close numerical approximation to it. (3) A transformation of it into a Poisson probability provides an always-exact risk. (4) A scaled version of it is the centrepiece of the Nelson-Aalen estimator.

In the light of (2) and (3), and the wide access to the exponential function, it makes sense to *always* convert the expected number to a risk.

Increasingly, reports include risk curves rather than survival curves, showing them at plots that go 'up, not down.'[19]. In addition to reducing the wasted white space, this practice removes an oxymoron in the phrase 'cumulative survival often used to label the y-axis in survival curves: in such curves, the estimated proportions/percentages 'still *in* the *initial* state' must go *down*; it is the transitions (*from* the initial state) that are cumulated!

Software packages differ in their use of terms for 'going up' curves, leaving it unclear whether plotted is the number of events in a dynamic population of size 1, or the probability – derived from it – that expresses the risk.

The Appendix describes an analogy between an ID function applied to a cohort, and a compound penalty rate applied to a sum of money.

Appendix: A heuristic from modern-day 'interest' rates

Suppose the balance $S(t') = \$100$ you planned to leave untouched in an account is below the minimum at which the bank pays interest; instead, from time t' onwards, it *penalizes* you by *decrementing* from the account. Suppose also that, contrary to the usual practice, it does so in real-time – 'continuously,' i.e., every nanosecond, rather than weekly or daily or hourly – and uses 'compound' rather than 'simple' decrements. Thus, the decrement 'at' (the nanosecond corresponding to) time t is the *product* of the penalty rate, and the balance, $S(t)$, 'at' t. In the simplest case, the 'penalty rate' is a constant, say an 'annualized' rate of 18 '%' (for every million '$-years' in such accounts, the bank takes $180,000) , or it might vary with the financial market – as an effectively continuous function.

Purists will, of course, note at least two differences: this is a deterministic formula, with no probabilities (beyond the volatility of the market); and, in theory, if not in practice, the 100 dollars are infinitely divisible, whereas the 100 persons are not. Nevertheless, the $S(t)$ does behave like a 'survival' function, and the key is how it is linked to the initial amount via a *summary of the ID or hazard function.* Just what form the summary takes becomes evident if one considers the following questions: (i) how is the balance at time t'' related to the *penalty-rate function* over the interval (t', t'')? (ii) for calculation purposes, is it sufficient to know just the *average* of the rate function over the interval (t', t'') and the duration $(t'' - t')$?

References

1. Armitage P, Berry G, Matthews JNS. A textbook of Medical Statistics. 4th Edition, Blackwell Publishing, 2002. p568-583.
2. Ayas NT, Barger LK, Cade BE et al. Extended Work Duration and the Risk of Self-reported Percutaneous Injuries in Interns. JAMA. 2006; 296: 1055-1062.
3. Schröder FH, Hugosson J, Roobol MJ, et al. ERSPC Investigators. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009 Mar 26;360(13):1320-8. Epub 2009 Mar 18.
4. Ridker, P. et al. New England Journal of Medicine, Nov 20, 2008 Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated CRP.
5. Miettinen, O. S. (1976) Estimability and estimation in case-referent studies. Am. J. Epidem., 103, 226-235.
6. Vandenbroucke JP, Pearce, N. Incidence rates in dynamic populations. International Journal of Epidemiology, Volume 41, Issue 5, 1472-1479.
7. Hanley JA. An interview with Olli Mettinen. Epidemiology 2012. Available online at `http://bcooltv.mcgill.ca/ListRecordings.aspx?CourseID=6485`.
8. Farr W. On Prognosis . British Medical Almanack 1838; Supplement 199-216) Part 1 (pages 199-208) Edited by GB Hill, with introductory note by A Morabia, reprinted in Soz.- Präventivmed. 48 (2003) 219224.
9. Vandenbroucke JP. On the rediscovery of a distinction. American Journal of Epidemiology (1985) 121. 627-628.
10. Farr W. Vital statistics: a memorial volume of selections from the reports and writings of William Farr. Humphreys NA, ed. Offices of the Sanitary Institute, London 1886.
11. Chiang CL. The Life Table and Its Applications. Robert E. Krieger Publishing Company, Malabar, Florida, 1984.
12. Chiang CL. Introduction to Stochastic Processes in Biostatistics. New York, John Wiley & Sons, Inc, 1968, chapter 12.
13. Miettinen, O.S. 1985. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine.*
14. Morgenstern H, Kleinbaum DG, Kupper LL. Measures of Disease Incidence Used in Epidemiologic Research. International Journal of Epidemiology 1980, 9: 97-104.
15. Rothman KJ. Modern epidemiology 1986 Little Brown Boston.
16. Edmonds, T. R. (1832) The Discovery of a Numerical Law regulating the Existence of Every Human Being illustrated by a New Theory of the Causes producing Health and Longevity. London: Duncan. Available as an on-line digital version at `http://books.google.com`.
17. Linder A. Daniel Bernoulli and J. H. Lambert on Mortality Statistics. Journal of the Royal Statistical Society, Vol. 99, No. 1. 1936, pp. 138-141)
18. Rothman KJ, Greenland S, and Lash TL Modern Epidemiology. Lippincott, Williams and Wilkins; Philadelphia, 2008. Chapter 3.
19. Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Lippincott, Williams and Wilkins; Philadelphia, 1998. Chapter 3.
20. Rothman KJ. Epidemiology: a introduction. Oxford University Press. 2nd Edition. 2012.
21. Collett, D. (2003) Modelling Survival Data in Medical Research, 2nd edn. Boca Raton: Chapman and Hall-CRC.
22. Pocock SJ, Clayton TC, Altman DG. Survival plots in clinical trials: good practice & pitfalls. Lancet 2002;359:1686-1689.

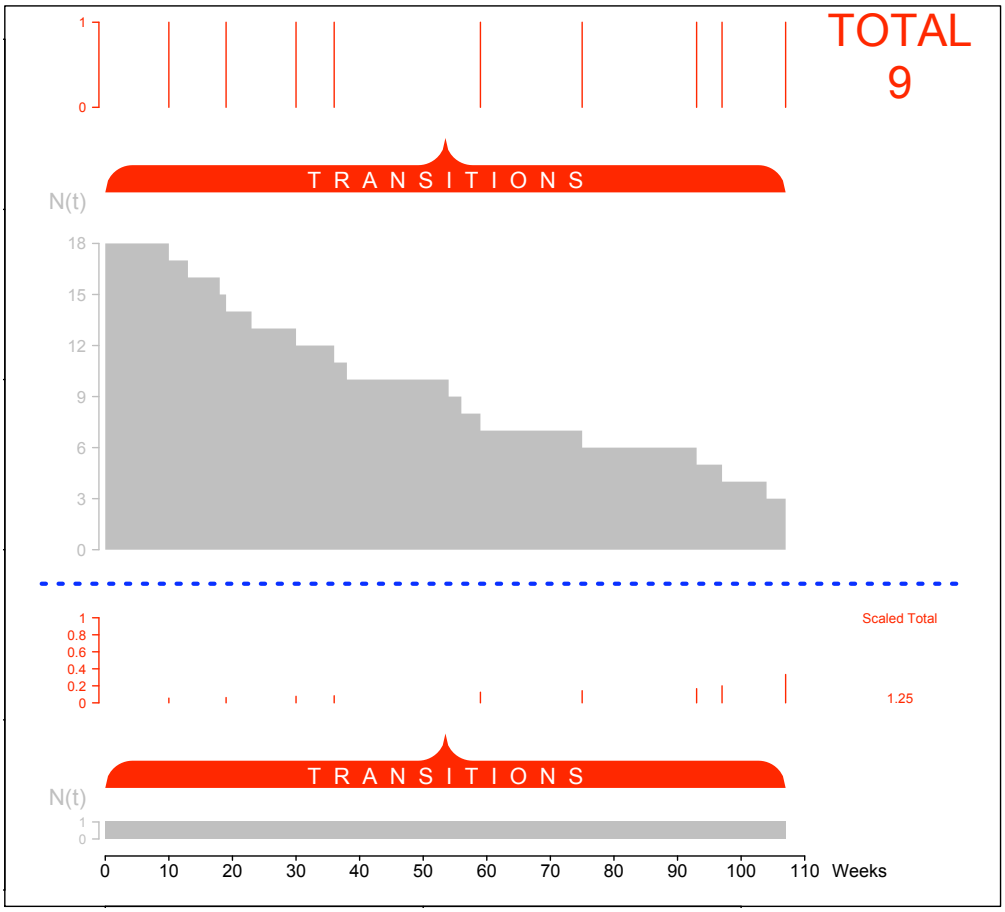

Figure 5: Heuristics for Nelson-Aalen estimator, using data on IUD discontinuation because of bleeding [21,p5]. 18 women began using an intrauterine device (IUD) for contraception, and were followed until the end of the study (entry was staggered) or until they discontinued it for unrelated reasons (total: 9 instances, treated as censored onservations), or until they discontinued it because of bleeding (9 instances). The upper panel shows the actual population-time using the function $N(t)$, i.e., the number of candidates at time t, and the timing of the 9 transitions. The lower panel shows the population-time scaled so as to always have one candidate, and the numbers of transitions scaled accordingly. Using the incidence density pattern in the top panel, we would expect $\mu = \sum_1^9 (d_j/n_j) = 1.25$ transitions in a dynamic population of size 1 followed for 107 weeks. Thus, the probability that a person who begins using an IUD at $t = 0$ will have discontinued it by $t = 107$ is $1 - \exp[\mu] = 1 - \exp[-1.25] = 0.71$, or 71%.