From incidence function to cumulative-incidence-rate / risk.
part II

Draft aug 15, 2012

James A Hanley
Department of Epidemiology, Biostatistics and Occupational Health
McGill University, Montreal, Canada.

possible outlets: Epidemiology, American J. of Epidemiology, ...

.

Comments welcome

Abstract

Although the terms mortality, hazard rate, incidence (rate), and incidence density involve the same concepts, those that involve continuous functions and mathematical limits, make many epidemiologists uncomfortable. Few textbooks present, and fewer still fully explain, the "exponential" formula linking incidence and risk. Better understanding of this link, and its underpinnings, is all the more critical today, as the familiar Kaplan-Meier estimate of a cumulative incidence proportion or risk is gradually being being replaced by the Nelson-Aalen one, and as investigators use parametric statistical models to calculate profile-specific x-year risks, risk differences, and numbers needed to treat, and to test proportional hazards via log[survival] plots. In part I, we illustrated the concepts common to the force of mortality, the hazard function, and incidence density functions We revisited the 1832 definition of the force of mortality and how a person-year was conceptualized, and used a striking 2010 graph to re-emphasize the centrality of time. With part I as orientation, we now extend the 1832 conceptualization, and use the probability of a specific realization of a Poisson random variate, to de-mystify the formula linking an incidence function and risk. We suggest ways to reduce confusion caused by variations in terminology.

1 Introduction and outline

The terms mortality, hazard rate, incidence (rate), and incidence density all involve the same concepts, but those that involve a mathematical limit (derivative) or integral make many epidemiologists uncomfortable. Indeed, although epidemiologists are comfortable with the concept of full-time equivalents in measuring staff sizes, this comfort level does not always extend to the concept of an intern-month or intern-year, or to converting an incidence function to a cumulative incidence proportion or risk. As a result, epidemiologists may be unsure as to how to turn an injury rate of say 0.095 needle-stick injuries per intern-month into a 12-month cumulative incidence or risk, and of what assumptions are involved. Indeed, few textbooks present, and fewer still explain, the formula linking incidence and risk.

Better understanding of this link is all the more critical nowadays, as the familiar Kaplan-Meier estimate of risk is gradually being being replaced by the Nelson-Aalen one, and as investigators use non-parametric and parametric statistical models to calculate profile-specific x-year risks (Schröder 2009), risk differences, and numbers needed to treat (Ridker, 2008).

In part I, the first of this pair of articles, we illustrated the concepts common to the force of mortality, the hazard function, and incidence density functions We revisited the 1832 definition of the force of mortality and how a person-year was conceptualized, and used a striking 2010 graph to re-emphasize the centrality of time.

With part I as orientation, this second of the pair addresses our main objective – demysifying the formula used to convert an incidence function to a cumulative incidence rate. To do so, we first review how the 'exponential' formula linking incidence and risk has been presented in epidemiology textbooks and articles. We then take advantage of Edmonds' conceptualization of a person year as "one person 'constantly living' for one year" (what today would be termed a dynamic population of constant size 1). We couple it with a little-used property of the Poisson distribution to de-mystify a 200 year old formula that seems to have been presented in a complicated way in modern textbooks. We illustrate how easy and unforgettable it is once its single input is fully understood, and how it is also the basis for the Nelson-Aalen estimator of survival or risk. We end with some recommendations about terminology.

2 Concepts and Historical background

2.1 Definitions

Incidence density refers to the rate of transition from a specific initial state (usually, but not necessarily, a health state) to a different specific state of interest. This rate is typically a function of age or time. As we did using the USA data from 2000-2006, it can be estimated from a *dynamic popula-*

tion experience[1], in which "a population of a given size but with turnover of membership moves over *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, it can be estimated from a *cohort* experience, "in which an enumerable set of individuals, all candidates initially, moves over the risk period."

The *cumulative incidence rate* is a *proportion-type* rate. It refers to a cohort (fictional or real) – all members of which are candidates initially – for a specified period or span of time [or age]. It is the *proportion which, in the absence of attrition, makes the transition in that period.* When the proportion is used as the *probability* of transition for an *individual*, it is usually referred to as a *risk*. This distinction between an expected proportion in an aggregate, and the probability for an individual is usually credited to Miettinen [1976, p229], but is also very clear in the writing of Farr [1838, p2], who 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.'

Typical applications are the 30-day mortality rate, the 1-, 5- and x-year risks of various illnesses, etc. A cumulative incidence rate can refer to an expected proportion in the abstract (theoretical, and thus a 'parameter' in statistical language), or to an empirical quantity.

[1]Some of the wording in this section is adapted from that in Miettinen 1976 and Miettinen 1985.

When a cohort experience is available (or envisioned), and each member has been followed up to the event at issue or to the end of the risk period, the cumulative incidence rate can be directly calculated as the proportion of the population of candidates, defined as of some zero time ($T = t_0$), who experience the transition during the risk period at issue. If there is attrition due to loss to follow up or extraneous mortality, the proportion can be calculated as the complement of the Kaplan-Meier or Nelson-Aalen survival function evaluated at the end of the risk period at issue.

But what if we wish to calculate the 20-year risk of death for persons aged 39.25, using the USA data from 2000-2007? Since this source population of subjects is dynamic – with new people continually entering at the lower bound (and within the range) of each age-interval and others exiting it (within the range and) at the upper bound, and a maximal membership duration of 7 years – it is not possible to directly calculate the proportion of the population of candidates, defined as of $t_0 = 39.25$, who *would* die during the 20-year risk period at issue. However, it is possible to do so indirectly using the statistical methods used to make 'current' lifetables. In this synthetic approach, the data from successive age categories (say 1 year wide) are 'spliced together' to project the experience of the hypothetical cohort. If data are abundant, the curve formed by joining the 'l_x's – the projected percentages still alive at the end of each year– by straight lines will be relatively smooth. But what if we had fewer data, and wish to calculate a smooth survival curve [$S(t)$] from a smooth incidence density curve $ID(t)$, such as the one displayed in

Figure 2(B)? Or what if we wish to convert an incidence density of 0.0975 (first) percutaneous injuries per month —assumed *constant* over a 12-month risk period, into a 12-month cumulative incidence (proportion-type) rate or risk?

2.2 The 'exponential' formula

Chiang (1984, p198) tells us that 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".

The coverage of this equation in the modern epidemiology era begins with Miettienen 1976. His 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." Since our example addresses the 20 year risk of death from *any* cause for 39.25, 59.25 and 79.25 year old persons, competing risks are not relevant. Thus, the formula given by Miettinen can be used without qualification: the cumulative incidence-rate (CIR) for the age span a' to a'' is (in his notation, but with his ID_a changed to $ID(a)$),

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

Miettinen gave, without commentary, the source for this equation as Chiang

(1968).

In 1980, Morgenstern et al. explain that if one assumes a fixed cohort and a constant death rate over a given interval , then with 'a little calculus', one can show that that constant rate and and the risk over that interval are mathematically related by this same exponential formula, which can be extended to cover risks over several periods, each with its own constant rate.

In his 1985 textbook, Miettinen again describes "the direct [algebraic] relation between incidence density (ID) and [the conceptual] cohort (cumulative) incidence (CI).

> Specifically, incidence density determines for a cohort (defined at $T = t_0$) the proportion which *in the absence of attrition* experiences the event before some common, quantitatively defined subsequent point in the time ($T = t_1$). With ID_t the ID at $T = t$, the CI for the interval t_0 to t_1 is (Chiang, 1968, Miettinen 1976a)
>
> $$CI_{t_0,t_1} = 1 - \exp\Big[- \int_{t_0}^{t_1} (ID_t) dt \Big].$$

As he had done in 1976, he also gave the version where the integral is replaced by a finite sum, but provided no insight into the 'anatomy' of either the continuous or the discrete version.

In 1985, Vandenbroucke wondered why the 150-year-old distinction between risk and rate had gotten lost, given that "with only slight alterations, excerpts from the mentioned texts by Milne and Farr would pass largely unnoticed in any modern textbook of epidemiology, if it were not for their exceptionally clear use of the English language." Vandenbroucke also refers to Farr's (p465) "formula for the calculation of the probability of dying from the rate of mortality and vice versa." However, the formula in question,

$CI_{x,x+1} = 1 - \frac{1-ID/2}{1+ID/2}$, still used today to calculate the 1-year risks (the $p_{x,x+1}$'s) for 'current' or 'period' population lifetables, is neither exact nor general.

2.3 Derivations/heuristics

Rothman (1986, pp 29-31) defines *cumulative incidence*, as "the proportion of a fixed population that becomes diseased in a stated period of time." He tells us that "it is possible to derive estimates of cumulative incidence from incidence rate." – again with the proviso that "there are no competing risks of death," and provides the mathematical formula that links cumulative incidence with the integral of the incidence rate function. Several epidemiologic textbooks since then have provided this mathematical expression, However, of the 15 modern texts JH has examined, only Rothman's 1986 textbook mathematically derives the relationship. Unfortunately, the formal geometric and calculus-based derivation it uses[2] does not provide any insight into 'why' or 'how' the 'exp' function comes into it. Thus, to may epidemiolo-

[2]the same one – with $S(t)$ as the solution of a differential equation – typically used in survival analysis textbooks, and also in 1980 by Morgenstern et al. This same approach was used by Edmonds (1832, p xvii) and, as Lidner (1936) recounts, (implicitly) by Lambert (1765) and Bernoulli(1776). The formula is similar to that required to answer the following question. The \$100 you leave untouched in a bank account at t' is below the bank's minimum at which it *pays interest*, so instead the bank *penalizes* you (in real-time – effectively continuously – rather than weekly or daily or hourly), by an amount that is applied to the balance, i.e. each decrement is the product of the penalty rate and the balance. In the simplest case, the 'penalty rate' might be constant, say an 'annualized' rate of 18 '%' (for every 100 '\$-years' in such deposits, the bank takes \$18) , or it might vary with the market – as an effectively continuous function. How is the balance at time t'' related to the penalty-rate function over the interval (t', t'')?

gists, especially in the absence of any worked examples, it remains a purely mathematical result.

Rothman's introductory textbook (2002, pp 33-38) uses heuristic arguments, but does not show the full-blown formula. Instead, it uses two worked examples. One assumed a mortality rate (incidence density) that remains constant – at 11 deaths per 1000 P-Y – over a 20-year age span, and, by proceeding year by year, as in a life-table, produced a cumulative incidence or risk of 19.7%.[3] The other addressed the risk, from birth through age 85, of dying from a motor-vehicle injury, assuming no competing causes of death, and 'piecewise-constant' rates of 4.7, 35.9, 20.1, 18.4 and 21.7 deaths per 100,000 person-years in the 5 age spans $0 \rightarrow 15 \rightarrow 25 \rightarrow 45 \rightarrow 65 \rightarrow 85$. The product of the 5 interval-specific conditional survival probabilities yielded an 85-year survival probability of 0.984 and thus a 85-year risk of 1.6%

In each example, the textbook used "the simplest formula to convert an incidence rate to a risk"

$$\text{Risk} = \text{Incidence rate} \times \text{Time}$$

However, it offered the following cautionary remarks [italics added] :

> It is a good habit when applying an equation such as [this] to check the dimensionality of each expression and make certain that both sides of the equation are equivalent. In this case, risk is measured as a proportion and has no dimensions. Although risk applies for a specific penod of time, the time period is a descriptor for the risk but not part of the measure itself. Risk has no units of time or any other quantity built in, but is interpreted

[3]The 20 year-by-year calculations in the first example (Table 3.2) would not have been any more complicated had the mortality rate changed from year to year rather than assumed to remain constant.

> as a probability. The right side of [the] equation is the product of two quantities, one of which is measured in units of the reciprocal of time and the other of which is simply time itself. *This product has no dimensionality* either, so the equation holds as far as dimensionality is concerned.

The text also urges end-users to check the *range* of the measures. Risk is "a pure number in the range [0,1]"; the product of incidence rate and time (both of which have "a range of $[0,\infty]$) can exceed 1." Thus, "the [above] equation is not applicable throughout the entire range of values for incidence rate and time," it is merely "an approximation that works well as long as the risk calculated on the left is less than about 20%."

We second these comments on units. However, rather than present an approach in which the product of ID and time is sometimes 'close to the numerical value of risk' and sometimes not, we prefer to explain that the *product has the same meaning no matter whether it is large or small, and that a simple transformation of it will always turn it into a risk (proportion).*

Chapter 3 in the 2nd and 3rd editions of Modern Epidemiology (1998, 2008) gives the discrete (i.e., summation) version of this 200-year old formula and tells us that it is sometimes referred to as the *exponential formula.* It is illustrated using a small numerical example, in which the Kaplan-Meier estimator yields a 19-year risk of 0.56. while the exponential estimator yields a risk of 0.52, leaving the reader to wonder which is an approximation to which.

We now give the product of ID and time (or more generally, the sum of products, i.e. the integral) in this 200-year old 'exponential formula' a

concrete meaning. This in turn will unveil the anatomy of the Nelson-Aalen estimator.

3 A different heuristic, inspired by Edmonds

To do so, we will take up Edmonds' concept of *a given number of persons constantly living*. Whereas he was concerned to keep the intervals small (in fact to use infinitesimal calculus) because he did not want the force to vary within the interval, ultimately we will consider much wider intervals, such as 20 years, where his assumption of a force *continued uniform* for that long – as is the one by Rothman2002 – would be unrealistic.

3.1 Less complex: constant-over-time ID

We begin with a simpler shorter-term example, in which we wish to convert an incidence density of 0.0975 (first) percutaneous injuries per month — assumed *constant*[4] over a span of 12 months – into a 12-month cumulative incidence (proportion-type) rate or risk.

As Edmonds did, we assume that the 'given number of interns' is one (1). We ask readers to imagine a 'chain', starting at $t' = 0$ and extending for 12 months until $t'' = 12$. The chain is begun with a randomly selected intern. That intern continues until he/she either reaches 12 months or is injured before then. If the latter, and if the intern is first injured at say

[4]Data from Ayas et al. 2006. We treat an intern-year as 3000 working hours, so that the ID= 0.00039 h^{-1}.

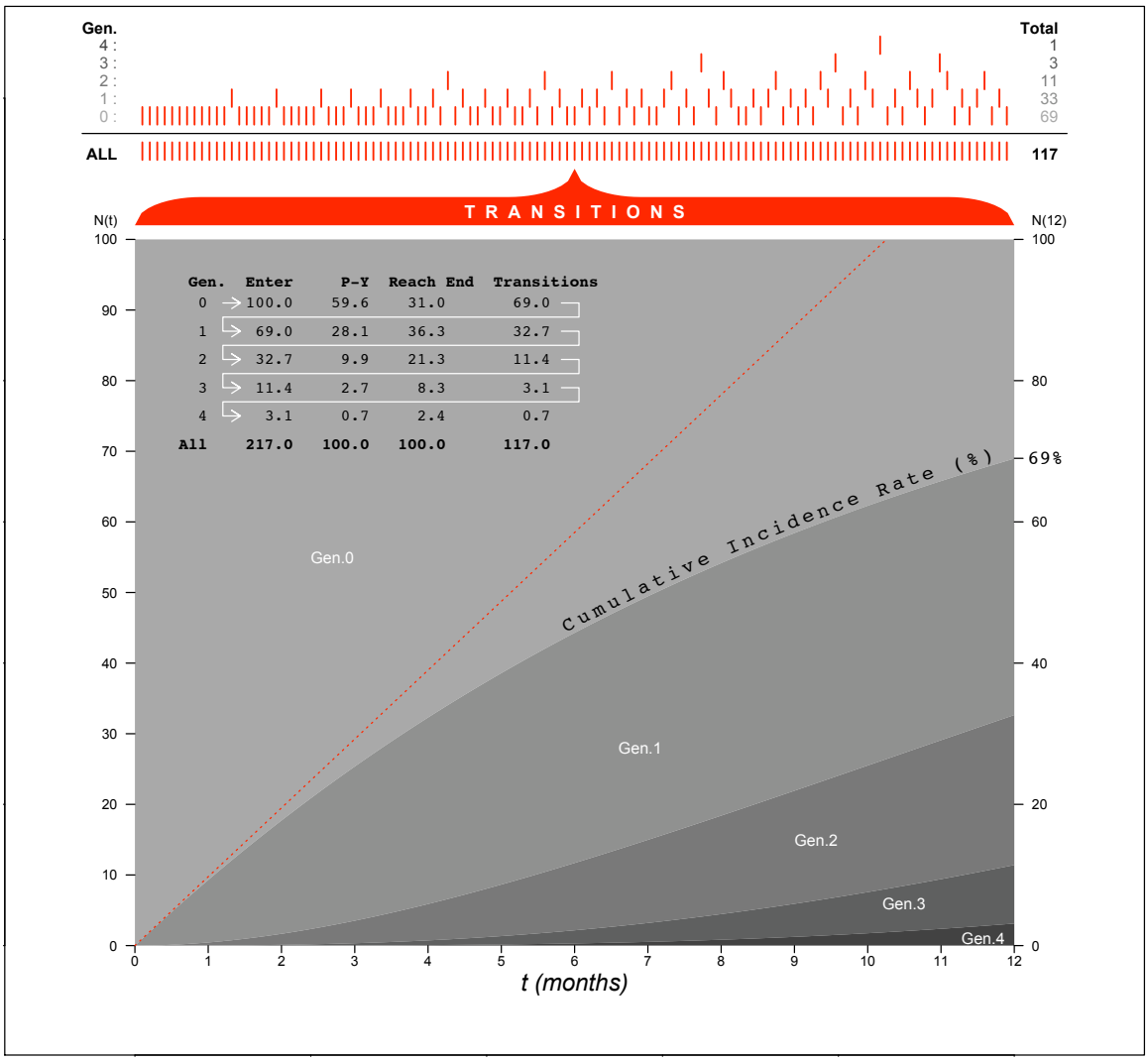

Figure 1: An average of 1.17 transitions (percutaneous injuries) in 1 intern-year (I-Y) of experience (117 in 100 I-Y), so that ID = 1.17 $year^{-1}$. 100 'chains' start at $t = 0$ (the 100 chains are represented by 100 horizontal lines, so close to each other that the total person time appears as a 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.

age t, he/she is immediately replaced by a a randomly selected never-injured intern. The chain proceeds, 'with further replacements as needed,' until it reaches $t'' = 12$. Throughout, there is 1 candidate, constituting a dynamic population with a constant membership of 1.[5]

The number of replacements required is a random variable, with possible values $0,\ 1,\ 2, \ldots$. 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.[6], we can first calculate the probability that the chain *is* completed by the same intern who initiated it as the Poisson probability of observing 0 events when 1.17 events are expected, i.e., as $\exp[-1.17] = \exp\big[-\int_{t'}^{t''} ID(t)dt\big] = 0.31$. The probability that the initial intern fails to complete the chain, i.e., *is injured before the 12 month period ends* is $1 - \exp\big[-\int_{t'}^{t''} ID(t)dt\big] = 1 - 0.31 = 0.69$. Thus the *12-month risk* of injury is 69%.

Fig 1, modeled after Fig 1 in Miettinen 1976, shows the expected values

[5]Another realistic 'chain' might be the experience, over a period $a'' - a'$, of a computer-server formed from a pool of exchangeable computers, all of the same age at time a': if the computer currently acting as the server fails, it is immediately replaced by another from the pool of computers still operating.

[6]Thus, it takes an average of 2.17 interns to provide the 1 intern-year of experience (in the computer- and other mission-critical examples, the years of experience – service – would be called 'up-time'.)

for a total of 100 separate such chains, and illustrates why *the product of ID and time (the 1.17, the integral) is not a risk per se, but rather an expected number of events (transitions, turnovers, injuries) in a dynamic population of size 1.* To accumulate 100 intern-years of service, an average of 217 interns is required. Of the 100 who initiated the chains (the average service of these 100, whom we might call '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 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.

Table 3.2 and Figure 3.3 of Rothman 2002 show a 20-year cumulative incidence rate, but using an incidence density of 0.011 yr^{-1}, so that the expected number of transitions in a dynamic population of 1 is $0.011\text{yr}^{-1} \times 1\text{ yr} = 0.22$. That 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

3.2 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 quantity 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 More complex: when ID varies over t

We deal 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 2(B). 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.[7] This sum – effectively the integral, from 79.25 to 99.25, of the ID

[7] This ('closed under addition') property of the Poisson distribution is well known to statisticians, but seldom exploited. Indeed, most epidemiologists – and many statisticians – insist that a Poisson random variate can only arise from single 'homogeneous' process. Yet, they – correctly – used the sum of observed numbers of cases over different age strata with very different incidence densities, as a Poisson random variate. In doing so, they are implicitly using the 'closed under addition' property of the Poisson distribution.

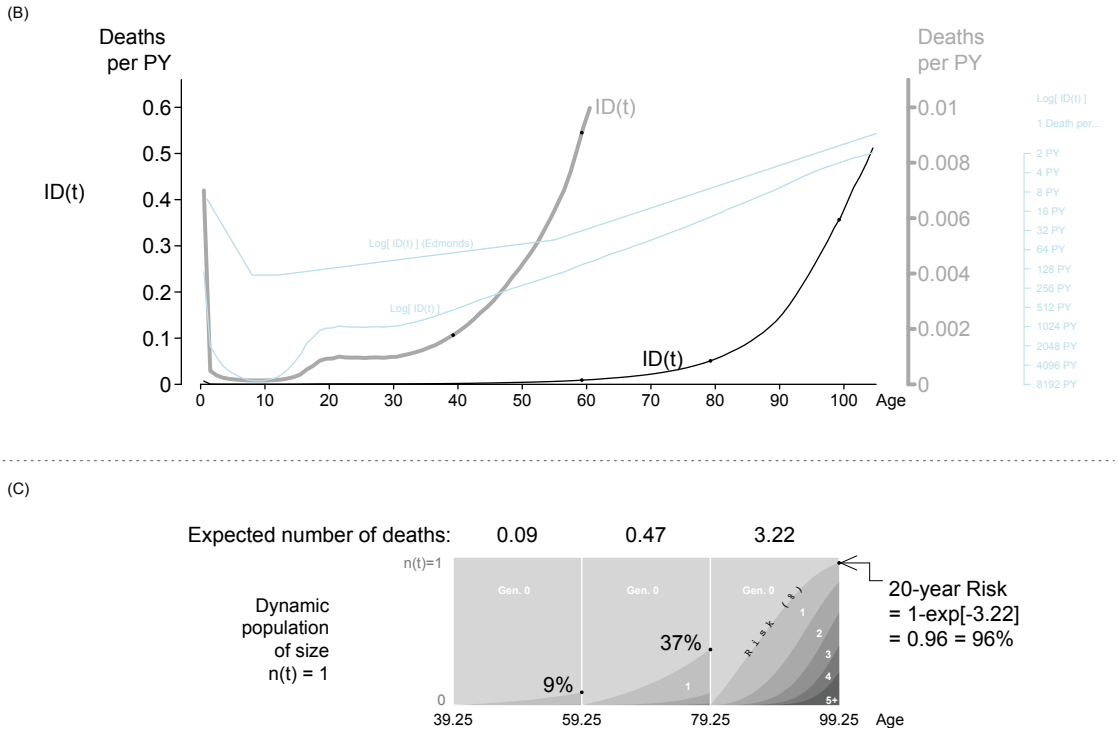

Figure 2: **(B)**, Age-specific death rates derived from age structure of, and age-specific numbers of deaths recorded in, the (dynamic) USA population observed over the period January 1, 2000 to December 31, 2006. For further details, and (A), see Figure 1 in part I. Shown are the full (in black, left axis) and 'below age 60' portion (grey, right axis) 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 $t = 105$. 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. **(C)**, 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 such spans. 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 $1 - \exp[-3.78] = 97.7\%$.

function in Figure 2(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 first 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 events 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 the complement of this, namely 1 - 0.04 = 0.96, or 96%. The risk curve (ie the risks for spans shorter than 20 years) is shown in the righmost panel of Fig 2(C).

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

In Rothman's 2002 example on the risk of dying of a motor-vehicle injury, the expected number of such 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 even fewer calculation steps that using the method he employed.[8]

[8] Although it is small enough to be a probability, the 0.016335 is not a probability *per se*. Rather, it is the expected number of deaths from injury if 1 person (not necessarily the same one) was constantly living'.

4 Approximation to CI

From the expected value of 0.09 in Figure 2(C), 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, are a reflection of the fact that, with a small expected value (E), so that $\exp[-E] \approx E$,

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

The $1 - \exp[-E]$ function can be closely approximated by E over the range $E = 0$ to $E = 0.1$, but this approximation becomes less accurate thereafter, as is shown by the following table[9]

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

The percentage over-estimation by using $\text{Risk}_{approx} = E$, rather than the exact expression $\text{Risk}_{exact} = 1 - \exp[-E]$, is close to $50 \times E$. Large values of E can arise from a low event rate operating over a longer time-interval, (e.g., 0.47 from mortality rates in the 20 year age span 59.25 to 79.25) or higher ones over a shorter one (e.g. 0.37 from mortality rates in the 1 year age span 99.25 to 100.25).

[9] Miettinen1976 merely states that "when the cumulative incidence-rate is small, say less than 10 per cent, it may be reasonably approximated by" this expected number; Rothman1986 explains: "because $e^x \approx 1 + x$ for $|x|$ less than about 0.1, it is a good approximation for a small cumulative incidence (less than 0.1). All of the textbooks that present the exponential formula caution about the limited range (some say $E \leq 0.1$, some $E \leq 0.2$) in which the approximation works.

As Vandenbroucke noted, Farr was 'aware that when the number of deaths is small, relative to the population studied, both measures approach each other numerically.'

5 The Nelson-Aalen estimator

The Nelson-Aalen estimator of the survival function (see Collett, 2003) has still to find its way into most epidemiology texts. It is usually presented as an 'alternative to' the Kaplan-Meier estimate. It is now included in most software packages and is increasingly found in the medical literature. It requires few mathematical operations than the Kaplan-Meier estimator. However, the most commonly presented heuristics – that the Kaplan-Meier estimator is an 'approximation to' the Nelson-Aalen one – do not give the full story, or explain why the Nelson-Aalen one is a natural estimator.

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 ignored.[10] 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

[10] Intervals with no events contribute multipliers of 1 to the product.

of the previous fraction, and so, ultimately, the estimator is simply the overall product of these:

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

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

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

Curiously, sometimes, it is justified by the statement that "the Kaplan-Meier Product Limit estimator is an approximation to it." This approximation holds true when *each* d_j/n_j is small, so that $1 - d_j/n_j \approx \exp[-d_j/n_j]$, and so that

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

But the Nelson-Aalen estimator of the survival function can also be thought of as the Poisson probability of 0 events when E are expected. This probability is exp[−E], where E is the number of events that would be expected if a certain $\widehat{ID}$ function i.e., a certain fitted force of morbidity/mortality function, were applied to a dynamic population with a constant membership of one ("one person constantly living"), over the time-span $(0, t)$. As above, $E = \int_{u=0}^{u=t} \widehat{ID(u)} du$. 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:

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

Fig 3 illustrates the heuristics using data on the frequency of IUD discontinuation because of bleeding (Collet, p5). The fitted number of transitions (discontinuations), $\sum_1^9 (d_j/n_j) = 1.25$, 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 $1 - \exp[-1.25] = 71\%$.

5.1 Terminology

The Nelson-Aalen estimator is increasingly used, but unfortunately, it has led to some confusion. This stems from the fact that the expected number of events in a 1-person dynamic population is sometimes close to the risk, and sometimes not, and that descriptions are not always clear as to which of these two numbers is being reported. Statisticians tend to refer to the expected number of events, i.e., the sum of products or integral, as the 'integrated hazard' or the 'cumulative hazard'. These terms should not confuse, but – as Rothman et. al (1998, 2008) lament – the term "cumulative inci-

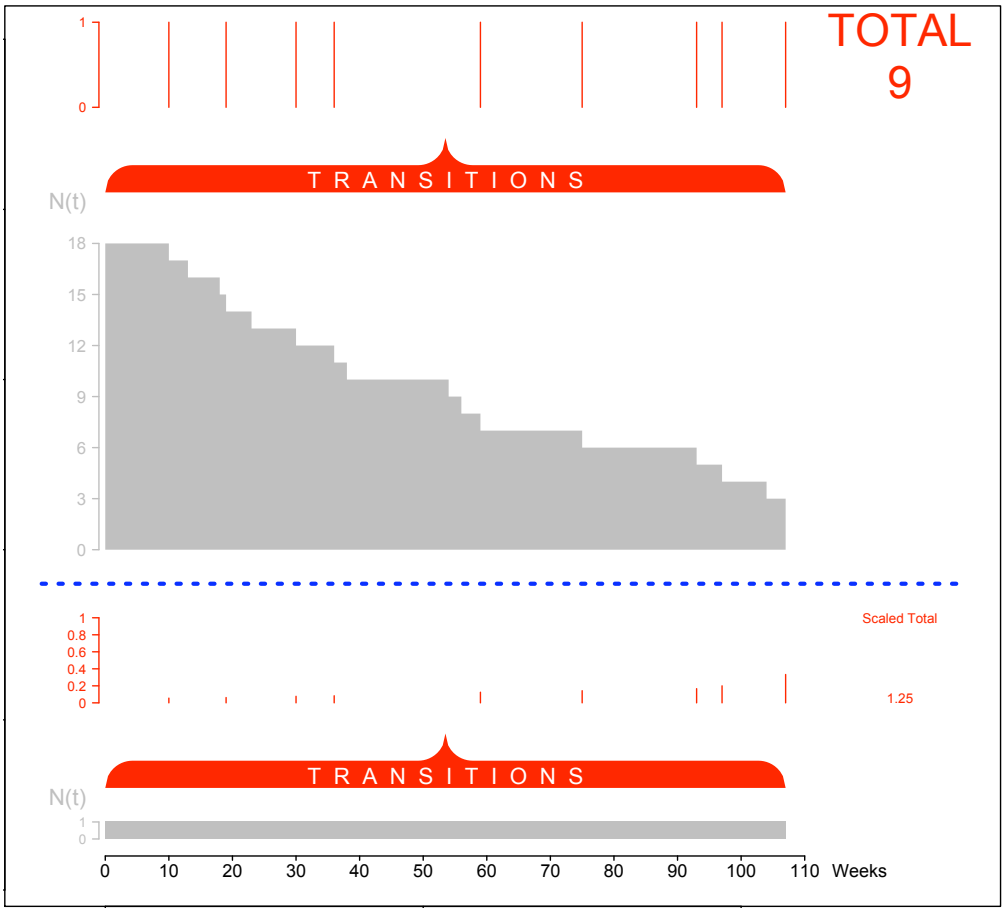

Figure 3: Heuristics for Nelson-Aalen estimator, using data on IUD discontinuation because of bleeding (Collet, p5). 18 women began using an intrauterine device (IUD) for contraception, and were followed until the end of the study (entry was staggered) or until they discontinued it for unrelated reasons (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 $\sum_1^9 (d_j/n_j) = 1.25$ 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 $1 - \exp[1.25] = 0.71$, or 71%.

dence" certainly could. To avoid just this possibility, throughout 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.[11] Stata software can calculate and plot the "the Nelson-Aalen cumulative hazard." As the user can verify using a dataset with a large expected number of cases (transitions) (e.g., the IUD one), what is indeed produced and plotted is an increasing (cumulative) set of expected numbers – each one a sum of products (an integral). Thus, they are not risks. But as Rothman 2002 and several others explain, the expected number will, in *low-expected number situations*, give a *reasonable approximation to the risk*. In such circumstances, the cumulative hazard will not greatly overstate the risk. However, it will do so when the expected number is high enough. Unfortunately, in the intermediate range where it is above say 0.1 but does not exceed unity, the user may not recognize that it is not a risk.

6 Recommended practice and terminology

So what should users do? First, we live in an age when everyone has ready access to the exponential function: it is even available on pocket calculators and smart phones. So, unless we are in extreme and unusual situations where we are forced to do the computations – division to get ID's, and multiplication

[11] I agree with Miettinen that epidemiologists do not have the right to proscribe use of the word rate to describe a proportion, when the word is widely used this way in common parlance; or to to restrict its use to a (time-based) transition rate.

and addition to get the expected numbers (integrals) – by hand, and cannot remember the series for $\exp[-x]$[12] we should *always* convert the expected numbers (the E's) into risks, using the *exact* formula $1 - \exp[-E]$. We have to compute E anyway, so the conversion to risk is only a small additional step.

Second, we should follow the advice of experts, and plot risk curves rather than survival curves (Pocock et al. 2002). They recommend should plots go 'up, not down.'

Third, if need be, we should either ourselves use the 'exponential equation' to convert the "the Nelson-Aalen" cumulative hazard values from Stata into risk values, or prevail on the Stata developers to make this an option.

Fourth, now that we know they are conceptually different – even if sometimes they have close to the same numerical value – we should not – as some have done – label the vertical axis the "Nelson-Aalen cumulative *hazard*" but entitle the figure the "Cumulative *Risk* of Death from Cancer."

Last, should we consider avoiding altogether the words cumulative *incidence*, or cumulative incidence *rate*, or cumulative incidence *proportion*, and instead simply use the word *risk*? I can think of two reasons to do so. One, it is the term used when referring to the output of 'risk-prediction' equations. There is no confusion when we see the words "Risk Assessment Tool for Estimating Your *10-year Risk* of Having a Heart Attack[13]. Two, even though

[12] $\exp[-x] = 1 - x + x^2/2 - x^3/6 \ldots$.
[13] http://hp2010.nhlbihin.net/atpiii/calculator.asp

both Farr and Miettinen have taught that (a) the cumulative incidence rate (or cumulative incidence proportion) is a *population* concept, and that (b) risk refers to the probability for an *individual*, in the end we use (a) as an estimate of (b). So, why not just use (b) directly and avoid (a)? Doing so might not be terminologically correct, but the amount of confusion that it would avoid might be worth it, and it would be unlikely to do much damage. It would also be good to reduce the use of the confusing word 'cumulative'. In a 't-year risk' curve, plotted against t, the word 'cumulative' is probably redundant. And in a 't-year cumulative survival' curve (a common default wording in software packages), the word 'cumulative' is an oxymoron – survival curves (the estimated proportions/percentages still in the *initial* state) go *down*; it is the transitions (*from* the initial state) that are cumulated!

References

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.

Chiang CL. Introduction to Stochastic Processes in Biostatistics. New York, John Wiley & Sons, Inc, 1968, chapter 12.

Chiang CL. The Life Table and Its Applications. Robert E. Krieger Publishing Company, Malabar, Florida, 1984.

Collett, D. (2003) Modelling Survival Data in Medical Research, 2nd edn. Boca Raton: Chapman and Hall-CRC.

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`.

Eyler JM. Constructing vital statistics: Thomas Rowe Edmonds and William Farr, 18351845. Soz.- Präventivmed. 47 (2002) 006-013, 2002

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.

Gompertz, B. (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining life contingencies. Phil. Trans. R. Soc. Lond., 115, 513-583.

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

Miettinen, O. S. (1976) Estimability and estimation in case-referent studies. Am. J. Epidem., 103, 226-235.

Miettinen, O.S. 1985. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine.*

Morgenstern H, Kleinbaum DG, Kupper LL. Measures of Disease Incidence Used in Epidemiologic Research. International Journal of Epidemiology 1980, 9: 97-104.

Pocock SJ, Clayton TC, Altman DG. Survival plots in clinical trials: good practice & pitfalls. Lancet 2002;359:1686-1689.

Ridker, P. et al. nejm nov 20, 2008 Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated CRP.

Rothman KJ. Modern epidemiology 1986 Little Brown Boston.

Rothman KJ. Epidemiology: a introduction. Oxford University Press. 2002

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Lippincott, Williams and Wilkins; Philadelphia, 1998.

Rothman KJ, Greenland S, and Lash TL Modern Epidemiology. Lippincott, Williams and Wilkins; Philadelphia, 2008.

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.

Turner EL and Hanley JA. Cultural imagery and statistical models of the force of mortality: Addison, Gompertz and Pearson. J. R. Statist. Soc. A

(2010) 173, Part 3, pp. 483-499.

Vandenbroucke JP. On the rediscovery of a distinction. American Journal of Epidemiology (1985) 121. 627-628.