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

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

Abstract

Although the terms mortality, hazard rate, incidence (rate), and incidence density all involve the same concepts, those that involve continuous functions and mathematical limits, and conversion of incidence functions to cumulative incidence rates, make many epidemiologists uncomfortable. Some of this has to do with the role of integrals and derivatives in the definition of these quantities, and in tasks such as converting an injury rate of say 0.095 percutaneous injuries per intern-month into a 12-month cumulative incidence or risk. Indeed, few textbooks present, and fewer still fully explain, the "exponential' formula linking incidence and risk. Increased understanding of this link is all the more critical nowadays, 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.

We use actual force of mortality (hazard function, incidence density) functions to illustrate how x-year risks are calculated from them. We revisit an early definition of the force of mortality – a term coined by an actuary colleague of William Farr – and describe how he viewed of a person-year. We take advantage of his conceptualization, and a 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.

0 Introduction and outline

Although the terms mortality, hazard rate, incidence (rate), and incidence density all involve the same concepts, 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.

Increased understanding of this links 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).

Section 4 addresses our main objective – demysifying the formula used to convert an incidence function to a cumulative incidence rate. By way or orientation, section 1 reviews the concept of a hazard or an incidence density in a single interval defined by a timepoint t. We use data from a dynamic population experience to measure the force of mortality (hazard function,

incidence density) over the life course, and comment on some 19th century attempts to fit it using smooth parametric functions.

Since the role of time, and time-units, in the measurement of the incidence function is often neglected, section 2 presents a striking time-graph to illustrate how critical time units are, and how easily they are overlooked, or even misunderstood.

Section 3 revisits the first definition of the force of mortality, a term coined by TR Edmonds, an actuary colleague of William Farr. Whereas most of us think of a person-year as the unit by which we measure an aggregated amount of experience of different persons, this actuary's concept of a person-year was somewhat different.

In section 4, we first review how the 'exponential' formula linking incidence and risk has been presented in various epidemiology textbooks, and why it needs to be further de-mystified. We then take advantage of Edmonds' conceptualization, and a little-used property of the Poisson distribution, to de-mystify this 200 year old formula. 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, in section 5, with some recommendations about terminology.

1 Incidence rate; incidence density; hazard rate; force of mortality

Epidemiology texts tend to define an incidence rate or incidence density using 'numbers of events' and person-time denominators, in intervals of time. Statistical texts define their equivalents, hazard rate or force of mortality, as mathematical limits involving probability distribution functions and survival functions measured in continuous time. The twain seldom meet.

The 2009 Wikipedia[1] entry for the 'instantaneous hazard rate', presumably prepared by a statistician, is a case in point. It defines it, in words first, as "the limit of number of events per unit time divided by number at risk as time interval decreases", and then in symbols:

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{observed events in interval } [t, t+\Delta t) \ / \ N(t)}{\Delta t}.$$

The entry did not define $N(t)$, but presumably it was used to denote the number at risk at time t, and allowed for the possibility that it might be different at time $t + \Delta t$, and at intermediate times. But if we consider a finite interval $[t, t + \Delta t)$, replace $N(t)$ by $\overline{N}$, the average no. persons being observed during the interval $(t, t+\Delta t)$,[2] and then move it to the denominator,

[1] The entry has since been edited.

[2] $\overline{N} \times \Delta t$ is also the integral of the $N(t)$ function over the interval in question.

then we get, for this finite interval, the quantity

$$\frac{\text{no. events in } (t, t+\Delta t)}{\overline{N} \times \Delta t} = \frac{\text{no. events in } (t, t+\Delta t)}{\text{Population-Time in } (t, t+\Delta t)}$$

which takes the form of incidence density, a term introduced to epidemiology by Miettinen (1976):

> 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.

In light of this, is the hazard rate or force of mortality the same as the 'short-term incidence density'? A concrete example in which we calculate the hazard rate or force of mortality at a given age t, based on the numbers of deaths, and person-years of observation in the USA population over the period 2000-2006, shows that it is, and that there is no need to be frightened by the mathematical limit, or to actually carry out the calculation using a tiny interval.

The 'almost-raw' data for our calculations, as well as the full ID function derived from it, are shown in Figure 1. The population sizes and numbers of deaths were only available for 1-year age bins; thus, in order to display them as a continuous function of age, the numbers were smoothed out using a spline function, so that the integral under the curve over any specific age-interval is the numbers of person-years lived in, or the numbers of deaths

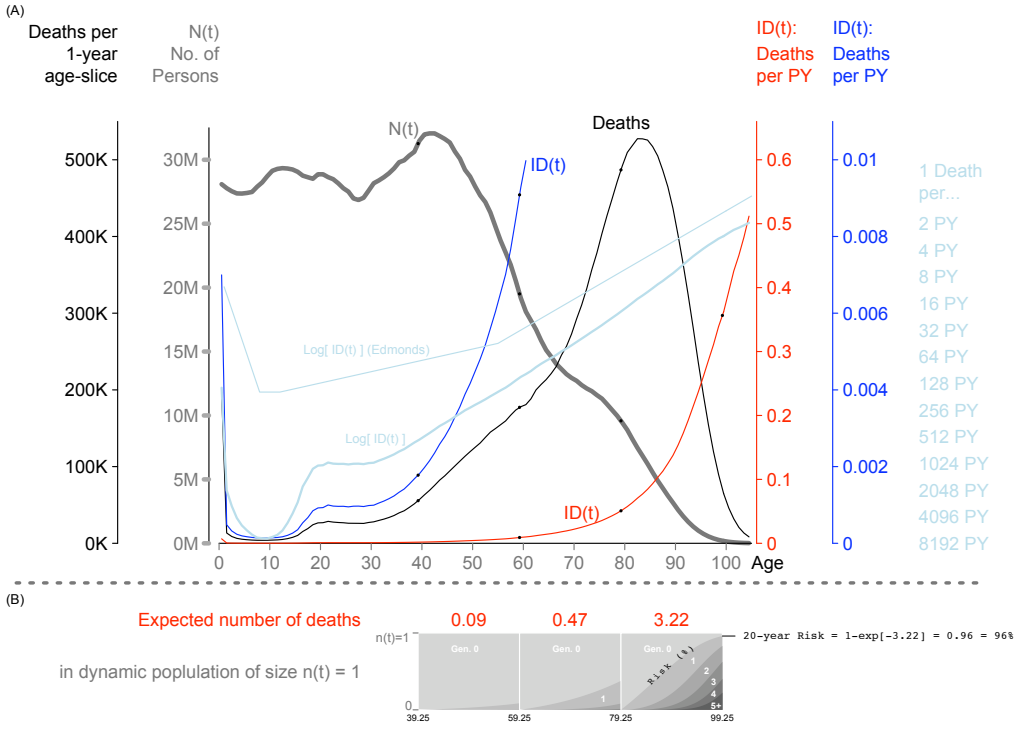

Figure 1: **(A)** Age structure of, and age-specific numbers of deaths recorded in, the USA population followed from January 1, 2000 to December 31, 2006, along with the death rates derived from them. Source: Human Mortality Database, `http://www.mortality.org/`. For each (continuous) value of age (t), shown 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 over the interval in question. The full $ID(t)$, or force of mortality or hazard rate function, ranges from a nadir of 0.000014 year^{-1} at approx. $t = 10$, to 0.51 year^{-1} at age $t = 105$. The $ID(t)$ function below age 60 is shown in a separate blown-up scale, and the log – to the base 2, so that we can easily measure doubling times – of the $ID(t)$ function is shown on yet another scale. Gompertz' Law of Mortality, in which the rate 'doubling time' is approximately constant (the logs of the rates are approximately linear) appears to hold true for the age range 30-90. For historical interest, Edmonds piecewise-linear log(ID) curve, based on data from early 1800s, is also shown on this scale. **(B)** 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" 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 there is at least one replacement, where 3.22 replacements are expected.

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

	t = age 39.25			t = age 59.25			t = age 79.25		
Δt	P-T*	Deaths	ID	P-T	Deaths	ID	P-T	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

*P-T Units: 1 million person-years ID Units: deaths / 100,000 years

Based on polulation-sizes and numbers of deaths, USA 2000-2006.
Source: Human Mortality Database, `http://www.mortality.org/`

recorded for, the age-interval at issue. Over the age span 0-105, there were approximately 17 million deaths in just over 2,000 million person-years.[3]

We take as illustration the ID or force of mortality or hazard rate at the (deliberately selected to be a non-integer) age $t = 39.25$. Technically, persons are only exactly 39.25 for a moment (infinitesimal, since a moment has no duration) and so we can only consider the calculation over say the finite interval $(t - \frac{\Delta t}{2},\ t + \frac{\Delta t}{2})$, of width Δt, that includes $t = 39.25$.[4] Table 1 shows the calculations with successively shorter intervals. The IDs would ultimately become unstable if we considered intervals as short as $39.25y \pm 3$ hours, say or $79.25y \pm 1$ hour. However, even as we narrow the intervals from a year to a day – or to minutes and seconds and nanoseconds if we ignore sampling

[3]Thus, the overall ID was 0.0085 year^{-1}; its reciprocal – 118 years – reflects the fact that this population experience is younger than in the current lifetable (expectation of life at birth: 77 years) calculated from these data.

[4]Since it doesn't fundamentally alter the concept, readers will find it easier, as we do here, to take t to be the *center*, rather than the *left boundary*, of the interval. The use of successively smaller intervals centered on a does cause some mathematical difficulties at $t = 0$, and explains why the limit is typically approached from the right.

variability and restrict our focus to the theoretical (i.e., abstract, expected) values – the ID's are practically unchanged. As Figure 1(A) shows, the ID(t) function does not change abruptly; it changes slowly and continuously.[5] Thus, the only reasons to be 'instantaneous' about it are if one wished to have a continuous smooth curve, especially one with a functional form, to shorten tedious annuity calculations or to compute an x-year risk (cumulative incidence), or to be able to provide an accurate break-even premium for 1-day term insurance for a large group of people.

We suspect that part of the 'divide' between statisticians and epidemiologists in this matter has to do with two different – but operationally equivalent – ways they define the hazard and the incidence density. Statisticians tend to first view it as a *theoretical* quantity and define it – in the abstract – as a (conditional) 'probability per unit time' for those who have reached t

$$\frac{\text{Prob[transition in next } \Delta t]}{\Delta t}$$

Indeed, Clayton and Hills (1993, chapter 5 (Rates), p40) give it a yet-another name:

> As the bands get shorter, the conditional probability that a subject fails during anyone band gets smaller. When a band shrinks

[5] over the 1-year interval centered on $t = 39.25$, the ID increases by about 0.021% per day or 8% in a year; for the 1-year interval centered on $t = 59.25$, the ID increases by about 0.024% per day or just over 9% over the year. These almost-constant year-over-year hazard ratios of 1.08 or 1.09 for much of the age-range are similar to those that Gompertz observed in the material he studied, and bear out the log-linearity of mortality rates with respect to age that he termed a Law of Mortality.

> towards a single moment of time, the conditional probability of failure during the band shrinks towards zero, but *the conditional probability of failure per unit time* converges to a quantity called *the probability rate*. This quantity is sometimes called the instantaneous probability rate to emphasize the fact that it refers to a moment in time. Other names are hazard rate and force of mortality.

Epidemiologists tend to first view it as an *empirical* quantity and define it using data. Indeed, Clayton and Hills *estimate* the rate parameter using the familiar incidence density measure:

> In general, then, as the bands shrink to zero, the most likely value of the rate parameter is
>
> $$\frac{\text{Total number of failures}}{\text{Total observation time}}.$$
>
> [...] This mathematical device of dividing the time scale into shorter and shorter bands is used frequently in this book, and we have found it useful to introduce the term *clicks* to describe these very short time bands. Time can be measured in any convenient units, so that a rate of 1.11 per year is the same as a rate of 11.1 per 10 years, and so on. The total observation time added over subjects is known in epidemiology as the person-time of observation and is most commonly expressed as person-years. Because

> of the way they are calculated, estimates of rates are often given the units *per person-year* or *per 1000 person-years*.

One way to reconcile the two is to recognize that Prob[transition in next Δt] is the expected number of transitions[6] as a fraction of the number of candidates. Thus, just as with the Wipipedia definition, when we divide this probability by Δt to get what Clayton and Hill call the probability of failure per unit time, it becomes

$$\frac{\text{No. transitions in next } \Delta t}{\text{Ave. no. candidates}} \div \Delta t \;=\; \frac{\text{Ave. no. transitions in next } \Delta t}{(\text{Ave. no. candidates}) \times \Delta t},$$

which has the the same form as Clayton and Hills' estimator.

Although statisticians and epidemiologists understand that "time can be measured in any convenient units, so that a rate of 1.11 per year is the same as a rate of 11.1 per 10 years, and so on," the next section shows that they sometimes forget how critical this point is, especially when one wishes to convert an (incidence-type) rate function into a risk.

[6]Since not all events in epidemiology involve movement from a more desirable (initial) state to a less desirable one, we use the more general term 'transition' instead of the term 'failure.'

2 The units in which incidence rate, incidence density, hazard rate, and force of mortality are measured

Figure 2 depicts the water demand time curves for (and presumably, the degree of exclusively-television-viewing by the residents of) the city of Edmonton the afternoon (and the afternoon before) the U.S.A. ice-hockey team played the Canadian team in the gold-medal game at the 2010 Vancouver Winter Olympic Games. The graph has been viewed by more people than has Minard's classic portrayal of the losses suffered by Napoleon's army in the Russian campaign of 1812.[7].

While the behavior pattern is striking, one omission from the label "Consumer Water Demand, ML" for the vertical axis is notable. After they realize that 'ML' is a measure of volume (it is short for 'megalitre' or millions of litres) aggregated over all consumers, people with engineering-type training, or physicians who measure lung function, whom this author has consulted have quickly responded that "it is missing a time-dimension." Curiously, epidemiology- and biostatistics-types have been slower to notice this omission. But even more interesting has been the split as to what they think the units of the missing dimension are. A number are quite confident that it must be 'ML per-hour' because "the time scale for the horizontal axis is marked

[7] `http://www.edwardtufte.com/tufte/posters`

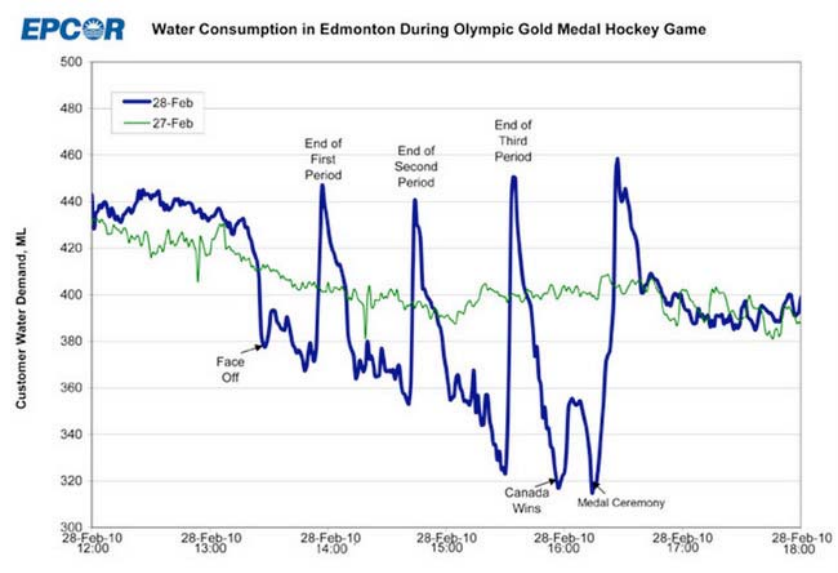

Figure 2: Minute-by-minute numbers of television-viewers of a major sports event. Q: What time units are missing from the label for the vertical axis?

off in hours". Others are equally adamant that it must be 'ML per-minute' because "the graph fluctuates by the minute." I leave it to readers to form their own opinions as to what the missing time unit is: those who like to calculate can use the following data: 400ML is approximately 106 million U.S. gallons; the population of Edmonton is approximately 700,000 people; "the average Edmonton resident uses 230 litres/person/day for indoor and outdoor use."[8]

To decide which unit most closely matches the reported usage, they will probably compute the total volume over the 6 hours, by taking the (approximate) integral of (area under) the water-demand curve over this time-span.

[8]http://www.epcor.ca/

But in order to do so, each Δt on the t-axis must be in the *same* units as the ML/timeunit on the demand-axis: the volume for each subinterval is $\frac{ML}{timeunit} \times \Delta t\ timeunits$. Thus, the total volume of demand for a given 15m interval is

$$Vol_{15min} =? \frac{ML}{hour} \times 0.25\ hours =? \frac{ML}{min} \times 15\ min =? \frac{ML}{day} \times 0.010416\dot{6}\ days$$

with $?\frac{ML}{timeunit}$ denoting the average demand per time-unit over the interval in question.

The reaction that the demand must be ML per minute because the time scale is in minutes is similar to the one which says that if we are to graph the velocity of a car over a period of minutes, we have to measure the speed in miles per minute rather than in mph– or that we cannot express heart rate in beats per minute if we only measure for 15 seconds. We can scale the velocity to any time unit we wish, but if the integral is to represent the total distance travelled, we need to calculate the distance travelled in each different subinterval of time as the (average) distance per time unit over that subinterval $\times$ the time-length of that subinterval – with the time-duration expressed in the same units as was the velocity.[9] This issue of time units

[9]For a striking example of improper use of units, and by the confusion caused by the statement that the "venous thromboembolic incidence was 3.6% and 1.5% in the first and second weeks postpartum, respectively, similar to the 2% to 5% incidence of symptomatic venous thromboembolism after elective hip replacement in patients not receiving prophylaxis" (when "only 105 maternal cases of venous thromboembolism were diagnosed during pregnancy or postpartum in 50 000 births"), see the correspondence regarding "Incidence of Pregnancy-Associated Venous Thromboembolism and family history as major risk factors" begun by MacCallum et al. in the Annals of Internal Medicine, 21 March 2006

becomes paramount in section 4, when we convert an incidence function into a risk.

3 Edmonds, the continuous force of mortality, and the concept of a person-moment and a person-year

Benjamin Gompertz used the word '*intensity*' of mortality in his 1825 article.[10] We believe that the first person to use the term '*force*' of mortality in writing was T.R. Edmonds, a political economist and actuary, and a neighbor and collaborator of William Farr (Eyler 2002; Turner and Hanley, 2010). Edmonds put the term in italics and in quotes in the first paragraph of his 1832 book. He begins his theoretical treatment with the words (emphasis ours)

> The force of mortality at any age is measured by the number of deaths in a given time, *out of a given number constantly living.* The given time has been here assumed to be one year, and the

Annals of Internal Medicine Volume 144(6). pp 453-460.

For an striking example of how different units can make an incidence function look larger or smaller, epidemiology students might wish to convert the army losses in Napoleon's Russian campaign into incidence densities, as a function of elapsed time (d $^{-1}$, or m^{-1}, or y^{-1}) or distance (mile^{-1}, or Km^{-1}, or $^{\circ}$long.$^{-1}$). The data are available at `http://www.math.yorku.ca/SCS/Gallery/re-minard.html`.

[10]Linder (1936) tells us that in 1765, "mathematician Johann Heinrich Lambert, (1728-1777) was the first to direct attention to what he calls 'Lebenskraft,' that is, the force of vitality or the reciprocal of the force of mortality".

given number living to be one person;

Whereas he defined the force of mortality as "the quantity of death in one year for a unit of life at the assumed age" he conceded that "the force is changing continually" and so he gives a more hypothetical definition "the quantity of death on a unit of life which *would* occur by the action of this force *continued uniform* for the space of one year. Edmonds employed infinitesimal calculus to use the "relation of Dying to Living for *large* intervals of age to deduce and interpolate the relation corresponding to *small* intervals of age":

> "[S]ince this relation for annual intervals is continually varying, it is manifest, that the same principles which have led to the conclusion, that the variation is continued and annual, must lead to the conclusion, that the variation is monthly, and also to the conclusion, that the variation is diurnal, and even momental.[11]
>
> It may be assumed, therefore, that all Tables of Mortality represent the relation of Dying to Living as changing continuously, - that this relation is never the same for any two successive instants of age. I have used the term 'force of mortality,' to denote this relation at any *definite moment* of age. It would evidently be improper to use this term to express the relation of Dying to Living in yearly intervals of age; for the force of mortality at the

[11]The word 'person-moment' is used in Miettinen OS. Etiologic research Needed revisions of concepts and principles Scand J Work Environ Health 1999;25 (6, special issue):484-490.

beginning, at the middle, and at the end of any year of age, are all different."

Edmonds found a very simple law to describe the continuous change in the force of mortality "during the succession of years and moments, measured from the birth of any individual." Starting at the nadir – an incidence density of 1 death / 160 years in the age span 8-12 – "it is expressible as three consecutive geometric series, or by the ordinates of three contiguous segments of three logarithmic curves. He presented the rate of increase or decrease of the force of mortality, in a given time, assumed to be one year" as a Table in numbers and in logs (I have added the logs to the base e and base 2):

In Numbers	In Logarithms			
	base 10*	base e	base 2	Period over which Constant presides.
0.6760830	- 0.1700	-0.3914	-0.5647**	Infancy (from birth to 8 years of age).
1.0299117	+ 0.0128	+0.0295	+0.0425**	Manhood (from 12 to 55 years of age).
1.0796923	+ 0.0333	+0.0767	+0.1106**	Old Age (from 55 to end of life).

* Used by Edmonds; **Doubling times 1/ 0.0425 $\approx$ 24 years, and 1/0.1106 $\approx$ 9 years, respectively

This is similar to the Law discovered by Gompertz.[12] The smooth parametric function for the force of mortality, and the mathematical relationship between the integral of this function and the cumulative incidence proportion, allowed actuaries such as Edmonds and Gompertz, and 'vital statisti-

[12]Gompertz' actuary colleagues never forgave Edmonds for not giving greater credit to Gompertz for "the honour of first discovering that some connection existed between the Tables of Mortality [the proportion alive at age x] and the algebraic expression a^{b^t}."

cians' such as Farr to make considerably shorten their annuity and lifetable calculations (Eyler; Turner and Hanley).

In the next section, we will exploit Edmond's idea of "one person 'constantly living' for one year" to derive the fundamental relationship between an incidence function and the cumulative incidence proportion (risk) that seems to have been neglected or made unnecessarily complicated in modern textbooks.

4 Link between ID functions and cumulative-incidence-rate / risk

4.1 Definition; formula; previous heuristics

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 is typically a function of age or time. As was done using the USA data from 2000-2006, it can be estimated from a *dynamic population* experience[13] , 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 indi-

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

viduals, 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*. Typical applications are the 30-day mortality rate, the 1-, 5- and x-year risks of various illnesses, etc.

When a cohort experience is available, 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 1? 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?

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, and Miettinen's calculations assumed 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 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)
>
> $$CI_{t_0,t_1} = 1 - \exp\left[- \int_{t_0}^{t_1} (ID_t)dt \right].$$

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 step-function version.

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[14] does not provide any insight into 'why' or 'how' the 'exp' function comes into it. Thus, to may epidemiologists, 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%.[15] 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

[14] the same one – with $S(t)$ as the solution of a differential equation – typically used in survival analysis textbooks.

[15] 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.

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 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. First, the Kaplan-Meier estimator is used to arrive at a 19-year risk of 0.56. The exponential estimator yields a risk of 0.52, but a reader may 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.

4.2 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.

4.2.1 Less complex: constant-in-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*[16] 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 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.[17]

The number of replacements required is a random variable, with possible values $0,\ 1,\ 2, \ldots$. Its expected value (mean) is $\mu = 0.0975\ m^{-1} \times 12\ m =$

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

[17]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.

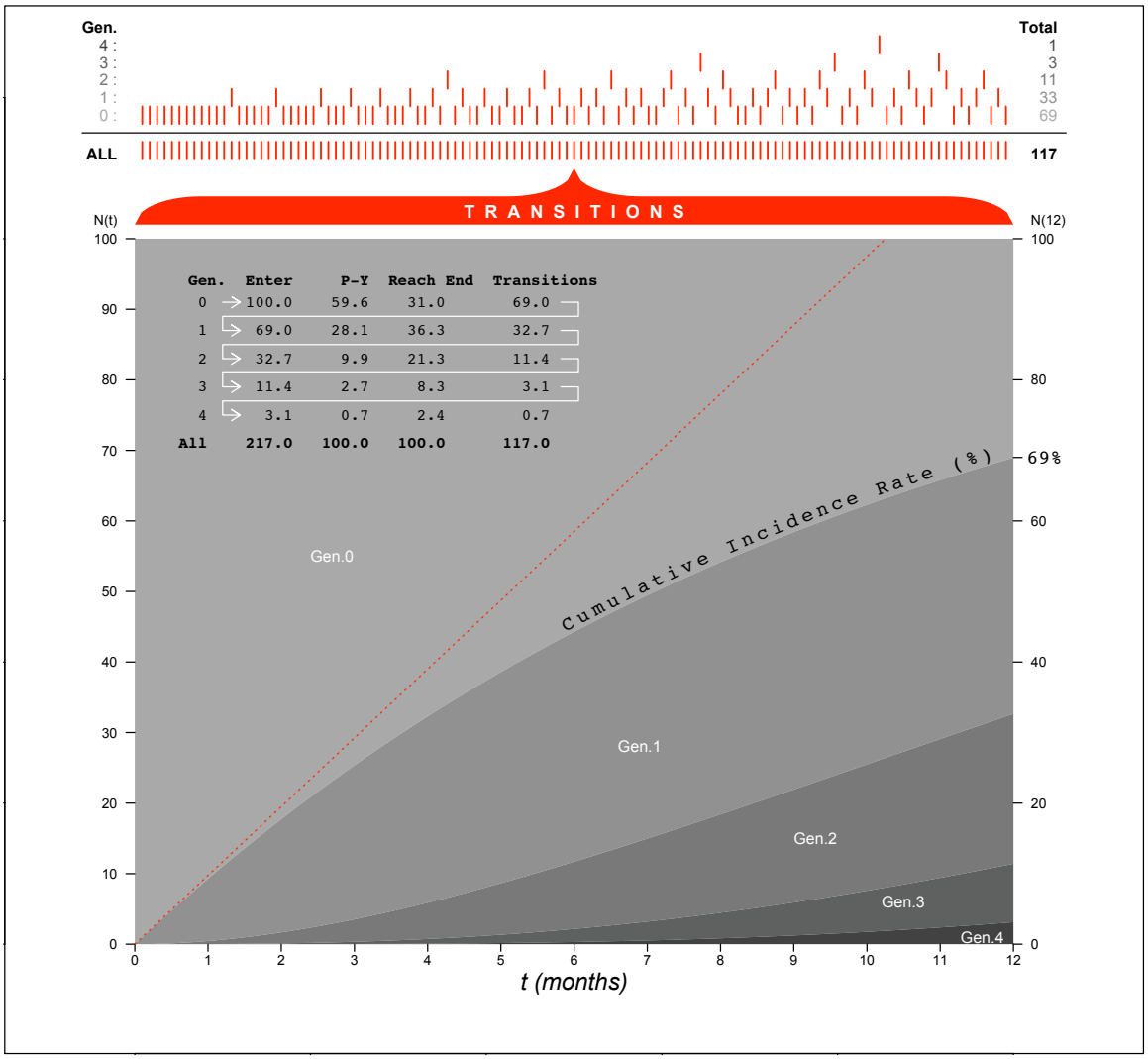

Figure 3: An average of 1.17 transitions (percutaneous injuries) in 1 intern-year (I-Y) of experience (117 in 100 I-Y), so that ID = 1.17 $year^{-1}$. 100 'chains' start at $t = 0$; each continues for 12 months, each using as many replacements (Gen. 1, 2, ...) as necessary to complete the chain. The 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.

$0.00039\ h^{-1} \times 3000\ h = 1.17$ 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.[18], 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 $\exp[-1.17] = \exp\left[-\int_{t'}^{t''} ID(t)dt\right] = 0.31$. The probability that the initial intern fails to complete the chain, i.e., *is injured before the 12 month period ends* is $1-\exp\left[-\int_{t'}^{t''} ID(t)dt\right] = 1-0.31 = 0.69$. Thus the *12-month risk* of injury is 69%.

Fig 3, modeled on Fig 1 in Miettinen 1976, shows the expected values for 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

[18]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'.)

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.

4.2.2 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 1(A). 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 $ID(t) \times (1/365.25)$. 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.[19] This sum – effectively the integral, from 79.25 to 99.25, of the ID function in Figure 1(A) – 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 full risk curve is shown in Fig 1(B).

[19] 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.

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.[20]

4.3 Approximation to CI

From the expected value of 0.09 in Figure 1, 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 $\exp[-E] \approx E$,

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

[20] 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'.

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[21]

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).

4.4 The Nelson-Aalen estimator

The Nelson-Aalen estimator of the survival function (see Collett, 2003) has still to find its way into 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

[21] 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.

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.[22] 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:

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

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

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

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, $E = \int_{u=0}^{u=t} \widehat{ID(u)} du$. This rectangular wave function 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:

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

Fig 4 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

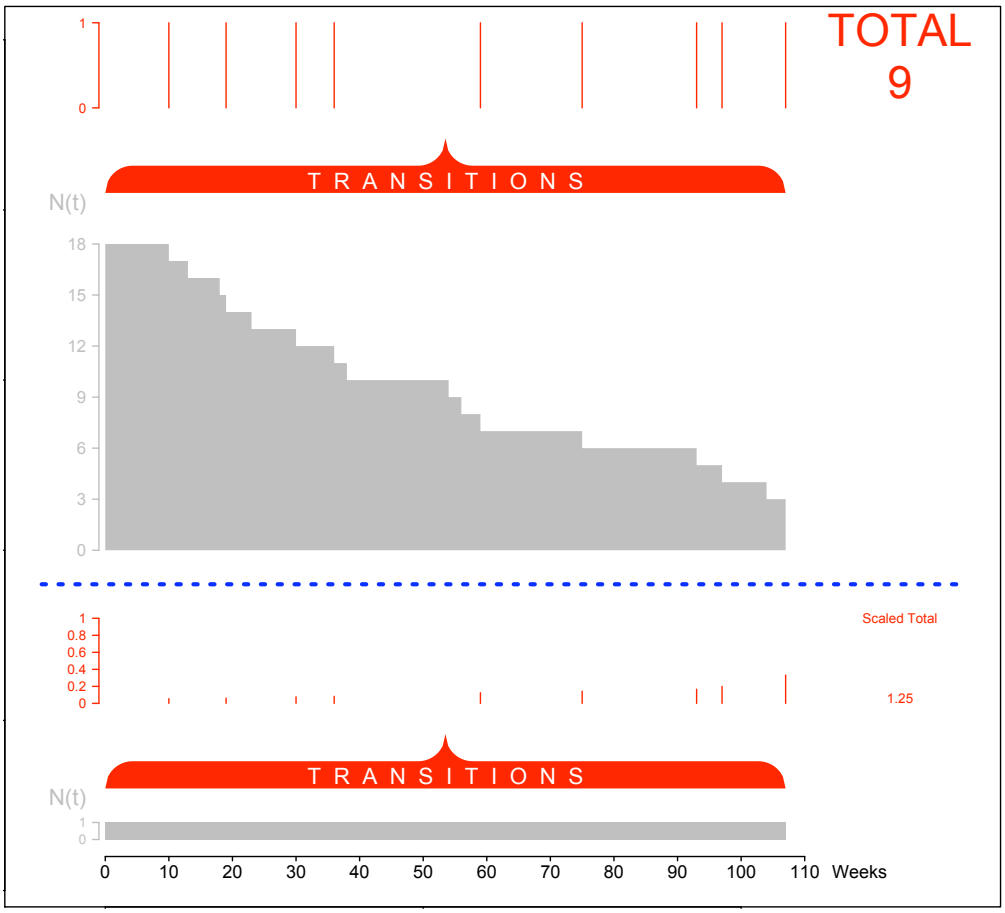

Figure 4: Heuristics for Nelson-Aalen Estimator, using data on IUD discontinuation because of bleeding (Collet, p5). 18 women began using an intrauterine device (IUD) for contraception, and were followed until the end of the study (entry was staggered) or until they discontinued it for unrelated reasons (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 $\sum_1^9 (d_j/n_j) = 1.246$ 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%.

always 1 candidate, and scaling the numbers of transitions accordingly. The 107-week risk is therefore $1 - \exp[-1.25] = 71\%$.

4.4.1 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 incidence" 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.[23] 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*,

[23] 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.

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.

5 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 and addition to get the expected numbers (integrals) – by hand, and cannot remember the series for $\exp[-x]$[24] 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

[24] $\exp[-x] = 1 - x + x^2/2 - x^3/6 \ldots$.

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[25]. Two, even though Miettinen teaches 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

[25]http://hp2010.nhlbihin.net/atpiii/calculator.asp

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.

Clayton D, Hills M. Statistical Models in Epidemiology. Oxford University Press, 1993.

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

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.*

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.