International Journal of Epidemiology, 2014, 949–961
doi: 10.1093/ije/dyu105
Advance Access Publication Date: 23 April 2014
Education corner

Education corner

Avoiding blunders involving 'immortal time'

James A Hanley[1] and Bethany J Foster[1,2]*

[1]Department of Epidemiology, Biostatistics, and Occupational Health and [2]Department of Pediatrics, Montreal Children's Hospital, Faculty of Medicine, McGill University, Montreal, QC, Canada

*Corresponding author. Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada. E-mail: james.hanley@mcgill.ca

Accepted 2 April 2014

As Groucho Marx once said 'Getting older is no problem. You just have to live long enough'.
(Queen Elizabeth II, at her 80th birthday celebration in 2006)

This award proves one thing: that if you stay in the business long enough and if you can get to be old enough, you get to be new again.
(George Burns, on receiving an Oscar, at age 80, in 1996)
(Richard Burton died, a nominee 6 times, but sans Oscar, at 59. Burns lived to 100, so how much of the 41 years' longevity difference should we credit to Burns' winning the Oscar?)

Some time ago, while conducting research on U.S. presidents, I noticed that those who became president at earlier ages tended to die younger. This informal observation led me to scattered sources that provided occasional empirical parallels and some possibilities for the theoretical underpinning of what I have come to call the precocity-longevity hypothesis. Simply stated, the hypothesis is that those who reach career peaks earlier tend to have shorter lives.
(Stewart JH McCann. *Personality and Social Psychology Bulletin* 2001;**27**:1429–39)

Statin use in type 2 diabetes mellitus is associated with a delay in starting insulin.
(Yee *et al. Diabet Med* 2004;**21**:962–67)

Introduction

For almost two centuries, teachers have warned against errors involving what is now called 'immortal time.' Despite the warnings, and many examples of how to proceed correctly, this type of blunder continues to be made in a widening range of investigations. In some instances, the consequences of the error are less serious, but in others the false evidence has been used to support theories for social inequalities; to promote greater use of pharmaceuticals, medical procedures and medical practices; and to minimize occupational hazards.

We use a recent example to introduce this error. We then discuss: (i) other names for it, how old it is and who tried to warn against it; (ii) how to recognize it, and why it continues to trap researchers; and (iii) some statistical ways of dealing with denominators measured in units of time rather than in numbers of persons.

Example and commentary

Example

Patients whose kidney transplants (allografts) have failed must return to long-term dialysis. But should the failed allograft be removed or left in? To learn whether its removal 'affects survival', researchers[1] used the US Renal Data System to study 'a large, representative cohort of [10 951] patients returning to dialysis after failed kidney transplant'. Some 1106, i.e. 32% of the 3451 in the allograft nephrectomy group, and 2679, i.e. 36% of the 7500 in the non-nephrectomy group, were identified as having died by the end of follow-up.

Patients in the two groups differed in many characteristics: to take into account a 'possible treatment selection bias', the authors constructed a propensity score for the

likelihood of receiving nephrectomy during the follow-up. They used this together with other potential confounders to perform 'multivariable extended Cox regression''. The main finding of these analyses was that 'receiving an allograft nephrectomy was associated with a 32% lower adjusted relative risk for all-cause death (adjusted hazard ratio 0.68; 95% confidence interval 0.63 to 0.74)'.

In their discussion, the researchers suggest that their findings of 'improved survival' after allograft nephrectomy 'challenge the traditional practice of retaining renal allografts after transplant failure'. The title of the article ('Transplant nephrectomy improves survival following a failed allograft') suggested causality. They emphasized the large representative sample and the extensive and sophisticated multivariable analyses, but they did caution that 'as an observational study of clinical practice, their analysis remains susceptible to the effects of residual confounding and treatment selection bias' and that 'their results should be viewed in light of these methodologic limitations inherent to registry studies'. They suggested that a randomized trial to evaluate the intervention in an unbiased way would be appropriate. Similar concerns about residual confounding and selection bias, and the need for caution, were expressed in the accompanying editorial reiterating the limitations of the 'retrospective interrogation of a database'.

Commentary

'Residual confounding' may be a threat, but both authors and editorialists overlooked a key aspect of the analysis, one that substantially distorted the comparison. The overlooked information is to be found in the statements that:

> 3451 received nephrectomy of the transplanted kidney during follow-up; the median time between return to dialysis [the time zero in the Cox regression] and nephrectomy was 1.66 yr (interquartile range 0.73 to 3.02 yr).
> (Paragraph 1 of Results section)

and that:

> Overall, the mean follow-up was (only) 2.93 ± 2.26 yr.
> (Paragraph 3 of Results section)

From these and other statements in the report it would appear that, in their analyses, follow-up of both 'groups' began at the time of return to dialysis. The use of this time-zero for the 3451 who had the failed allograph removed is not appropriate—or logical. These patients could not benefit from its removal until after it had been removed; but, as the median of 1.66 years indicates, a large portion of their 'follow-up' was spent in the initial 'failed graft still in place' state—along with those who never underwent nephrectomy of their failed allograft.

Since the 3451 patients who ultimately underwent a nephrectomy (the 'nephrectomy group') had to survive long enough to do so (collectively, approximately 6700 patient-years, based on the reported quartiles of 0.73, 1.66 and 3.02 years), there were, by definition, no deaths in these 6700 pre-nephrectomy patient-years. In modern parlance, these 6700 patient-years were 'immortal'. There was no corresponding 'immortality' requirement for entry into the 'non-nephrectomy group'. Indeed, all 10 951 patients returning to dialysis after failed kidney transplant began follow-up with their 'failed graft in place'. Some 7500 of these remained in that initial state until their death (for some, death occurred quite soon, before removal could even be contemplated) or the end of follow-up, whereas the other 3451 spent some of their follow-up time in that initial state and then changed to the 'failed graft no longer in place', i.e. post-nephrectomy, state.

How big a distortion could the misallocation of these 6700 patient-years produce? The article does not have sufficient information to re-create the analyses exactly. Figures 1 and 2 show a simpler hypothetical dataset which we constructed to match the reported summary statistics quite closely. It was created assuming no variation in mortality rates over years of follow-up or between those lived in the two states. The 'virtual' intervention was set up 'retroactively' and was limited to the dataset itself, rather than to real individuals, and so could not have affected (other than randomly) the mortality rates in the person-years lived in each state.

Figure 2A shows that even though the data were generated to produce the same mortality rate of 11.8 per 100 PY (person-years) in the person-years in the initial and post-'intervention' states, the inappropriate type of analysis used in the paper, applied to these hypothetical data, would have resulted in a much lower rate (6.4) in the 'intervention group and a much higher one (17.1) in the 'non-intervention' group. The reason is that none of the 1031 deaths post-'intervention' could have occurred, and none of them did occur, in the 6732 (immortal) pre-'intervention' PY that are included in the denominator input to the rate of 6.4: logically, the 1031 post-'intervention' deaths only occurred in the post-'intervention' PY. And conversely, the 2759 deaths occurred not in 16 096 PY, but rather in the much larger denominator of $16\,096 + 6732 = 22\,828$ PY lived in the initial state. The omission of the 6732 PY from the denominator input led to the rate, higher than it should have been, of 17.1 deaths/100 PY. Indeed it was because of these (misplaced) immortal 6732 PY they had already survived that the 3451 patients got to have the 'intervention'; in other words, it may not have been that they lived longer because they underwent the 'intervention', but rather that they underwent the 'intervention' because they survived

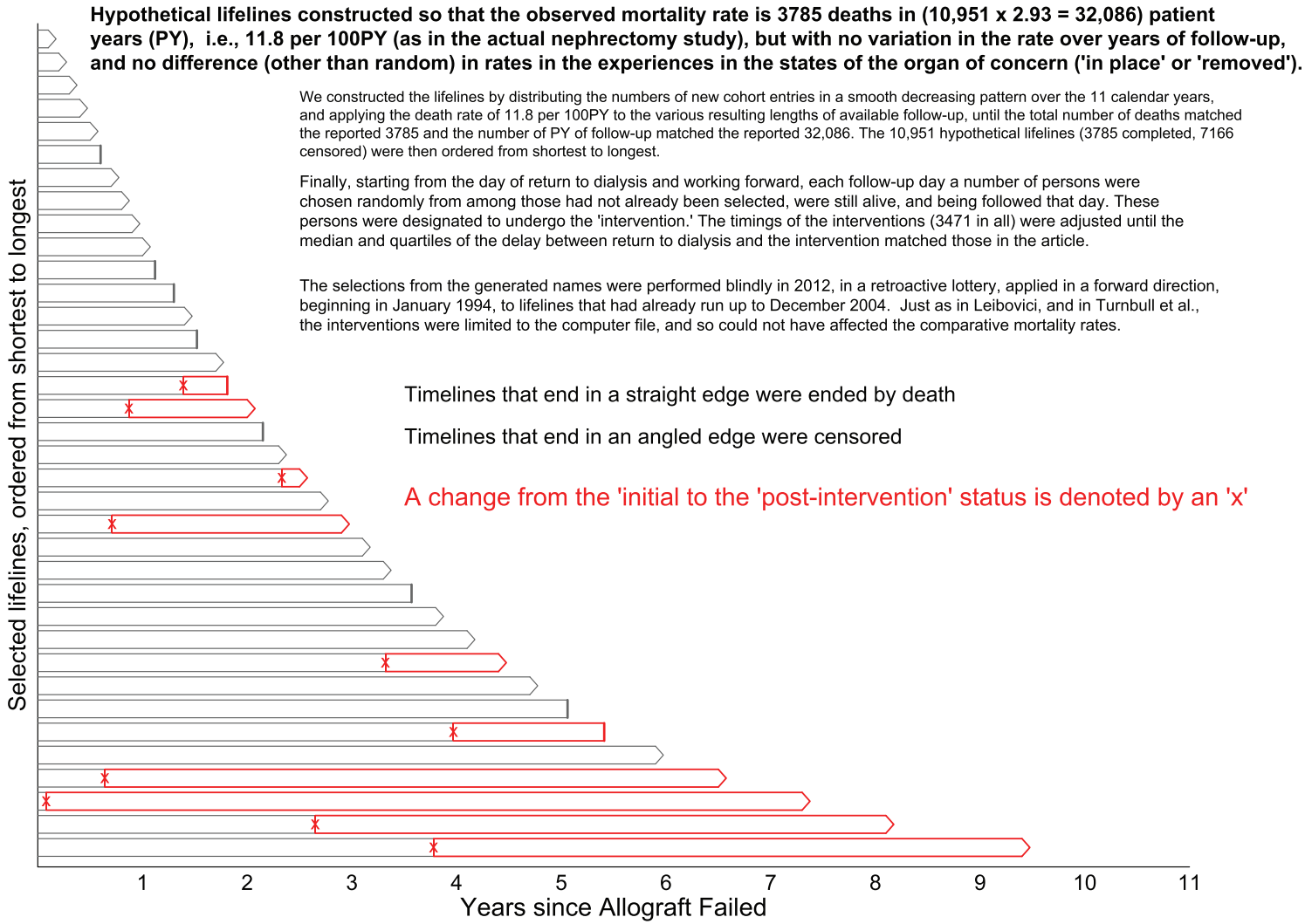

Figure 1. Excerpts from the simulated mortality experience in the contrasted ('organ intact' vs 'organ removed') states. Hypothetical lifelines were generated to have an average mortality rate of 3785 deaths in (10 951 × 2.93 = 32 086) patient-years (PY), i.e.,11.8 per 100 PY (as in the actual nephrectomy study[1]), but with no variation over years of follow-up, and no difference (other than random) between states ('name intact' or 'name removed'). We constructed the dataset by first generating names for 10 951 fictional persons, then distributing the numbers of new cohort entries in a smooth decreasing pattern over the 11 calendar years, and then applying the death rate of 11.8 per 100 PY to the various resulting lengths of available follow-up, until the total number of deaths matched the reported 3785 and the number of PY of follow-up matched the reported 32 086. The 10 951 hypothetical lifelines (3785 completed, 7166 censored) were then ordered from shortest to longest. Finally, starting from the day of return to dialysis and working forward, each follow-up day a number of persons were chosen randomly from among those who had not already been selected, were still alive and were being followed that day. These persons were designated to undergo an electronic 'removal' whereby, within the database, just their names (not their failed transplants) were (electronically rather than surgically) removed. The timings of these 'removals' (3471 in all) were set so that the median and quartiles of the delay between return to dialysis and becoming nameless matched the delays in the article. The selections, made by a random number generator in 2012, were made blindly, in a retroactive lottery, applied in a forward direction, beginning in January 1994, to lifelines that had already run up to December 2004. Just as in Leibovici[16] and in Turnbull *et al.*,[17] these interventions were limited to the 2012 computer file, and could not have affected the comparative mortality rates. Shown are 30 such lifelines selected systematically from these 10 951 ordered hypothetical ones, with a completed lifeline indicated by a single straight line, and a censored one by a pair of lines forming an arrowhead. The timing of the name removal is indicated by an x, and the post-intervention PY by red rather than grey boundary lines.

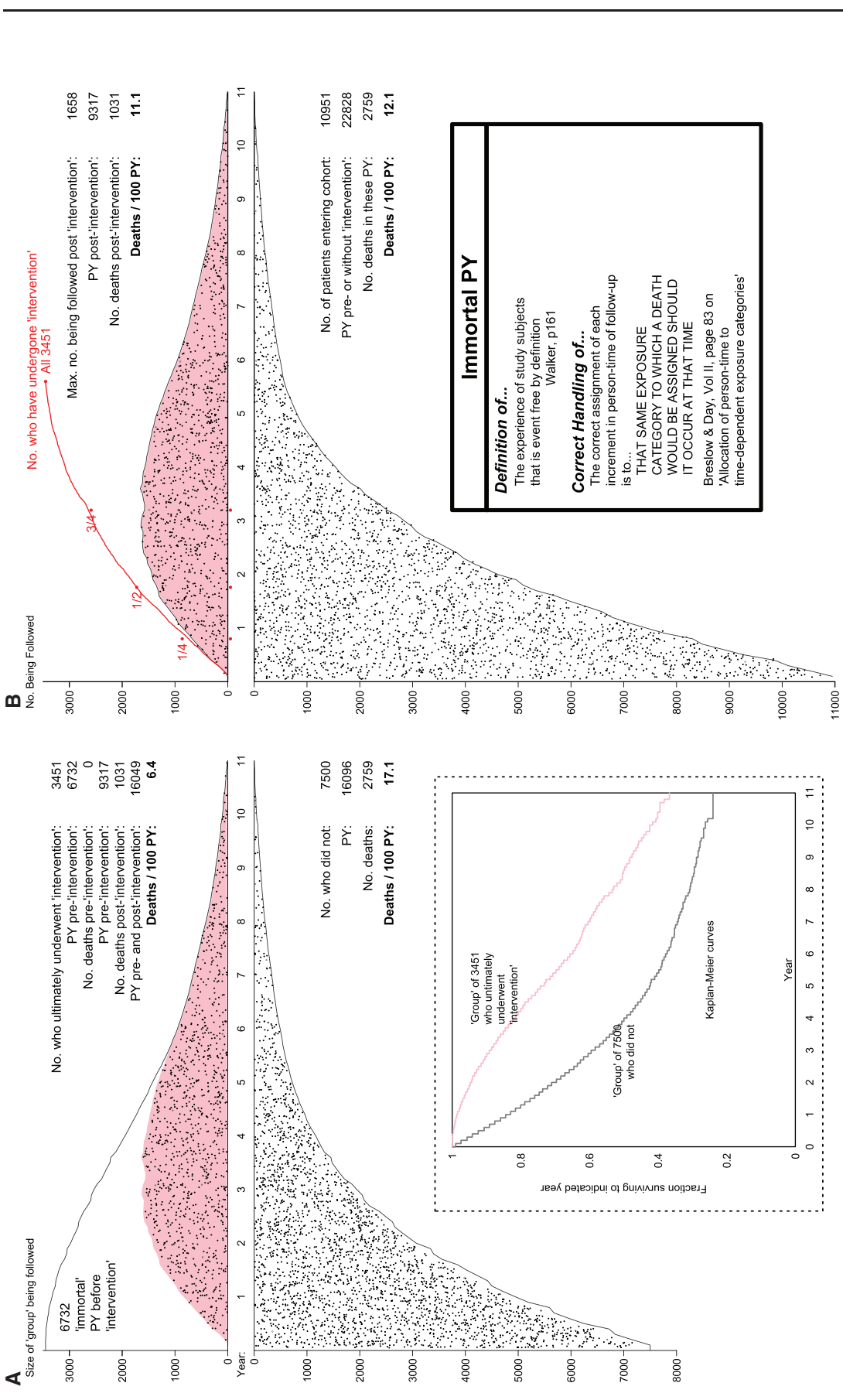

Figure 2. Mortality rates and rate ratios produced by the (A) mis- and (B) proper allocation of pre-'intervention' patient years. As explained in Figure 1, the hypothetical data for the 10 951 patients were constructed to have an average mortality rate of 3785 deaths in ($10\,951 \times 2.93 = 32\,086$) patient-years (PY), i.e. 11.8 deaths per 100PY (as in the actual study), but with no variation over years of follow-up, or between states (no, or pre-'intervention' (white background) and post-'intervention' (pink background). Indeed, the selection of those who changed states (from white to pink polygon, in B) was made at random, and retroactively. The time location (relative to when the allograft failed) of each death is indicated by a black dot. In B, upper panel, the number being followed at any time is smaller than 3451 because some who had received the 'intervention' were already dead before the last ones received it.

long enough to undergo it. One can see how, with some epidemiologists' penchant for long lists of biases, they might term this artefact 'reverse causality bias' (Senn, in a personal communication regarding reference 13, suggested that the 'higher mortality rates' in 'the childless' could equally be reported under the headline: 'Those who die young have fewer children'.)

In this admittedly over-simplified version of the data, with no covariates, the inappropriate analysis led to an apparent rate ratio of $6.4/17.1 = 0.37$. The corresponding 'reduction' of 63%, and an 'improved survival' of at least 2.5 years (areas under the first 11 years of the Kaplan–Meier curves of 7.7 vs 5.2 years), would have been interpreted as having been produced by the intervention, whereas they are merely artefacts of the misallocation of the PY.

Figure 2B shows an appropriate comparison of mortality rates in time-dependent states. With 'each unit of person-time allocated to the state in which the death would have been assigned should it occur at that time', the appropriate rates are (apart from random error) identical, mimicking the theoretical rates used to generate these hypothetical data. The theoretical rates were—unrealistically—taken as constant over the follow-up years. In reality, the PY in each year of follow-up time would be contributed by individuals who were almost one year older than the individuals who contributed PY the year before, and so the mortality rates in successive time-slices would also be successively higher. Thus, since the person-years in the post-'intervention' state are 'older' person-years, a summary rate ratio computed using matched slices of follow-up time would be more appropriate than a crude rate ratio. One would also need to match the person-years on several patient-related factors. As time-slices become more individualized, the distinction between Poisson regression, with its emphasis on the time interval, and Cox's approach, with its focus on the time moment, becomes more blurred. Space does not allow us cover these approaches here, but below (at the end of this article) we provide a link to some additional material we prepared on this topic.

Teachings against such blunders

Warnings against this error go back at least to the 1840s, when William Farr[2] reminded sanitarians and amateur epidemiologists that:

> Certain professions, stations, and ranks are only attained by persons advanced in years; … hence it requires no great amount of sagacity to perceive that 'the mean age at death', or the age at which the greatest number of deaths occurs, cannot be depended upon in investigating the influence of occupation, rank, and profession upon health and longevity.

Then, in an admirable style seldom equalled in today's writings, he explained that:

> If it were found, upon an inquiry into the health of the officers of the army on full pay, that the mean age at death of Cornets, Ensigns, and Second-Lieutenants was 22 years; of Lieutenants 29 years; of Captains 37 years; of Majors 44 years; of Lieutenant-Colonels 48 years; of general Officers, ages still further-advanced … and that the ages [at death] of Curates, Rectors, and Bishops; of Barristers of seven years' standing, leading Counsel and venerable Judges … differed to an equal or greater extent … a strong case may no doubt be made out on behalf of those young, but early-dying Cornets, Curates, and Juvenile Barristers, whose mean age at death was under 30! It would be almost necessary to make them Generals, Bishops, and Judges—for the sake of their health.

Crediting the years of immortality required to reach the rank that the person has reached by the time (s)he dies or follow-up ends exaggerates any longevity-extending benefits of reaching this rank. Likewise, crediting the time until one receives a medical intervention to the intervention exaggerates its life- or time-extending power.

Whereas Farr adopted a tongue-in -cheek style, Bradford Hill[3] spelled out the reason for the longevity difference: 'Few men become bishops before they have passed middle life, while curates may die at any age from their twenties upwards'. Separately,[4] Hill also addressed the fallacy under the heading 'Neglect of the period of exposure to risk':

> A further fallacy in the comparison of the experiences of inoculated and uninoculated persons lies in neglect of the time during which the individuals are exposed first in one group and then in the other. Suppose that in the area considered there were on Jan. 1st, 1936, 300 inoculated persons and 1000 uninoculated persons. The number of attacks are observed within these two groups over the calendar year and the annual attack-rates are compared. This is a valid comparison so long as the two groups were subject during the calendar year to no additions or withdrawals. But if, as often occurs in practice, persons are being inoculated during the year of observation, the comparison becomes invalid unless the point of time at which they enter the inoculated group is taken into account.

Hill used a worked example to warn that 'neglect of the durations of exposure to risk must lead to fallacious results and must favour the inoculated'. The example shows that the adjective 'immortal' time is not broad enough: 'event-free time, by definition or by construction' (see Walker, below[9]) is a more general and thus a more appropriate term.

Ten years earlier, Hill[5] had addressed the 'period of exposure to risk' when comparing, 'from age 25 to age 80', the longevity of cricketers with that of the general male population.

> The comparisons show that cricketers form by no means a short-lived population, but on the contrary hold a substantial advantage at every age … this advantage is undoubtedly somewhat exaggerated since it is assumed that all cricketers are 'exposed' from age 25, while in actual fact probably some do not 'enter exposure' in first-class cricket till a later age.

Breslow and Day[6] use a diagram, and a simplified occupational epidemiology example, modelled on the blunder by Duck *et al*.,[7] to emphasize the correct allocation of person-time, and the distortions produced by misallocation. In Figure 3 we illustrate the Duck *et al*. error and the reallocation of the person-years by Wagoner *et al*.;[8] our preference for vertical rather than horizontal shading is meant to illustrate the 'as you go' (vertical) rather than 'after the fact' (horizontal) accumulation of person-years. We also repeat Breslow and Day's succinct enunciation of the general principle to be followed.

We understand that the term 'immortal time' had been used by George Hutchison in the 1970s already, but his Harvard colleague Alexander Walker[9] is the first we know of to have put the term in writing, in his 1991 textbook. Walker's numerical examples all involve the correct allocation of such time, with no example given of the consequences of misallocation. The two editions of the Rothman and Greenland textbook[10] do have an example—albeit hypothetical—of the difference between incorrectly and correctly calculated rates based on two parallel groups of exposed and unexposed persons, and state the principle: 'If a study has a criterion for a minimum amount of time before a subject is eligible to be in the study, the time during which the eligibility criterion is being met should be excluded from the calculation of incidence rates'. They also allude, without an example, to the more general situation where subjects change exposure categories. Unfortunately, it is in this latter situation that most immortal-time blunders are made.

By using the term 'immortal time' in the title of a 2003 article, Suissa[11] immortalized the term itself. Since then, more than a dozen articles and letters by him and his pharmaco-epidemiology colleagues have addressed the growing number of serious 'immortal time' errors in this field. Typically, cohort membership in these studies was defined at the time of diagnosis with, or hospitalization for, a medical condition. The blunders were created by dividing the patients into those who were dispensed a pharmacological agent at some time during follow-up and those who were not (Unlike in most clinical trials, but like Hill's inoculation example, not all received it immediately at entry to the cohort.) When, instead, each patient's follow-up time is correctly divided into the portions where the event-rate of interest might be affected, and the portion where it cannot, the rate-lowering power of the agent disappears.

In several of their articles, Suissa and co-authors use other real datasets to address the same question, and show the consequences of the misallocation. Our annotated bibliography gives several other examples (and collections of examples), by yet other authors, of time-blunders in several other fields. However, even with warnings in one's own journals, time-blunders continue to occur: 1 year before it received the manuscript containing the study of transplant nephrectomy, the *Journal of the American Society of Nephrology* published an expository article[12] explaining how such a blunder can be recognized and avoided.

Recognizing and avoiding immortal-time blunders

Table 1 lists some ways to recognize immortal time and to avoid the associated traps. We suspect that some of the blunders stem from the tendency—no matter the design—to refer to 'groups', as though—in a parallel-arm trial—they were formed at entry and remained closed thereafter. Even when describing a cross-over trial, authors mistakenly refer to the treatment group and the placebo group, rather than to the time when the (same) patients were in the treatment or placebo conditions or states. This tendency may reflect the fact that many questions of prognosis can only be studied experimentally by parallel group designs. Except in studying the short-term effects of alcohol and cellphone use while driving, or medication use or inactivity on blood clots, cross-over designs (called split-plot designs in agriculture) are rare; and their statistical results are more difficult to show graphically and in tables than are those that use independent 'groups'.

Just as in the story of Solomon, it is appropriate that persons remain indivisible. However, in epidemiology many denominators involve amounts of time (yes, contributed by persons, but time nonetheless), and time is divisible, just as are any other (area- or volume-based) denominators that produce Poisson numerators (The numerators are not divisible.) Despite this, many epidemiologists are less comfortable with dividing an individual's time into exposed and unexposed portions than they are with measuring research staff size in full-time-equivalents, or than telephone companies are in measuring the amount of time used by customers. We look forward to the companies providing researchers with access to their information on the moment-by-moment location of users' cellphones,

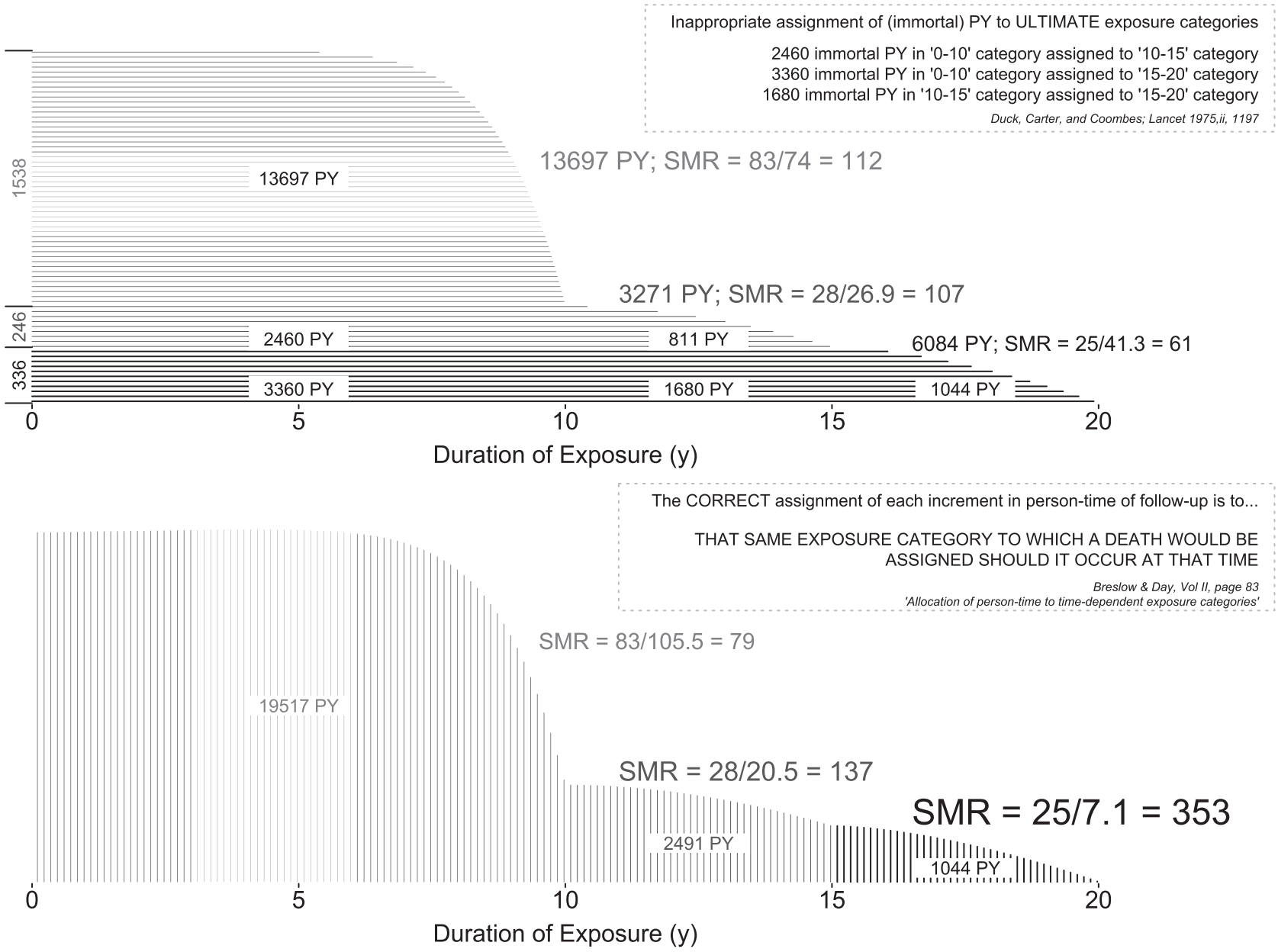

Figure 3. Incorrect (at termination) and correct (as time progresses) allocation of follow-up time in the Duck *et al.* study.[7] Horizontal timelines represent exposure to vinyl chloride, with durations categorized into 0–10, 10–15 and 15+ years. See also the small worked example, based on this study, in Breslow and Day.[6]

Table 1. Ways to recognize immortal time

Suggestion	Remarks/tests
Distinguish state from trait	A trait (e.g. blood group) is usually forever; people and objects move between states (on/off phone; intoxicated/not; on/off medication; failed allograft in place/removed)
Distinguish dynamic from closed population	Membership in a closed population (cohort) is initiated by an event (transition from a state) and is forever; in a dynamic population, it is for the duration of a state. Dynamic populations are the only option for studying transient exposures with rapid effects (e.g. cellphone/alcohol use vs the rate of motor vehicle accidents)
Focus on person-time in index and reference categories, rather than on people in exposed and unexposed 'groups'	These refer to exposure categories, not to people per se; a person's time may be divided between exposure categories; unless people remain in one category, it is misleading to refer to them as a 'group'
If authors used the term 'group', ask …	When and how did persons enter a 'group'? Does being in or moving to a group have a time-related requirement? Is the classification a fixed one based on the status at time zero, or later? Is it sufficient to classify a person just once, or do we need to classify the 'person-moments,' that is the person at different times?
Sketch individual timelines	If there are two time scales, a Lexis diagram can help; use different notation for the time portion of the timeline where the event-rate of interest might be affected, and the portion where it cannot (see Figures)
Measure the apparent longevity- or time-extending benefits of inert agents/interventions	After the fact, use a lottery to assign virtual (and never actually delivered) interventions, but with same timing as the one under study. Or use actually-received agents with same timing
Imagine this agent/intervention were being tested within a randomized trial	How, and when after entry, would the agent be assigned? Administered? How would event rates be computed? How would Farr have tested his 'early-promotion' suggestion?
Think short intervals and hazard rates, even if the hazard rates do not change abruptly	In addressing the present, conditional on the past, the hazard approachhas already correctly documented the experience in each small past interval; the natural left to right time-ordering of the short intervals allows for correct recognition of transitions between exposure states. By computing a mortality rate over a longer time-span defined after the fact, one may forget that in order to contribute time to the index category, people had to survive the period spent in the (initial) reference category

so that they can more accurately measure the amounts of on-the-phone and off-the-the phone driving time, and the rates of motor vehicle accidents in these. The more comfortable biomedical researchers become in dividing an individual's time (e.g. same person with different hearts), the less the risk of immortal time blunders.

It is our impression that epidemiologists are more comfortable 'splitting time' than many researchers in the social sciences, where correlations (rather than differences in and ratios of incidence rates) are the norm. If one is studying the duration of life, using lives that have been completed, it may not matter much whether one compares the aggregated lives divided by their number (average lifetime) or the total number (of deaths) divided by the total lifetime (the average number of deaths per unit of time). However, once one restricts attention to the rate of (terminating) events within just portions of these lifetimes, the switching between 'exposure' states, incomplete lives and censored and truncated observations all make it much more difficult to stay with the familiar correlations carried out using the time scale itself. The 'correlations between election age and death age for restricted subsamples based on election age percentile'[13] and the 'setting time-zero' to some arbitrary birthday (e.g. 0, 50 or 65 in the case of those nominated for an Oscar) are good examples of the limitations of staying with the average duration (longevity) scale that is easier to convey to the public. Those who compare rates (dimension: time^{-1}) within the relevant time-windows have much more flexibility than those who attempt to compare average durations (dimension: time).

Theories such as the just-cited precocity-longevity hypothesis are seductive, and have a certain plausibility. But some of this may be a result of the framing. A restatement of the 'evidence' can help uncover the fallacy: imagine if Groucho Marx were to re-word it, using Ronald Reagan's election and longevity as the example. In any case (as we stated in our re-examination of the claimed 3.9 year longevity advantage for Oscar winners see additional references), no matter how important or unimportant results would be if they were true, 'readers and commentators should be doubly cautious whenever they encounter statistical results that seem too extreme to be true'.

Failure to recognize immortal time errors leads to consequences that in some cases may be serious and costly,

and not easily corrected. In a case like the pharmacoepidemiology work of Suissa and colleagues, the costs of correcting the record can be considerable. Original stories in the lay press are not usually updated: the Oscar longevity story in the *Harvard Health Letter* of March 2006 is still available online—if one is willing to pay $5 for access. The Harvard website does not cite any updates, nor does the (itself widely cited) 2010 *New York Times* interview.[14] *Forbes Magazine*[15] is an exception. Medical journal editors, too, appear reluctant to correct blunders missed by peer review. Indeed, when one of us (B.F.) wrote to the Editor of the *American Journal of Nephrology* to point out the strong possibility that the finding of 'improved survival' following allograft nephrectomy was an artefact, she was told that the journal did not have a Letters to the Editor section, but that it would pass on the concerns to the authors.

Given HR Haldeman's observation, 'Once the toothpaste is out of the tube, it's hard to get it back in', it seems prudent and scientifically responsible to try to avoid immortal time errors from the outset. Researchers can avoid 'immortal time' errors by classifying person-time into exposure states, rather than classifying whole persons who ultimately attain an exposure state into 'exposed' and 'unexposed' groups, under the assumption that they have been in those groups from the outset.

Or, as Steve Jobs told us, 'Think different'. Think person-time, not person.

Additional material

Since the 'extended' Cox model is often used in this 'change of states' context, our first version of this manuscript contained a section entitled 'Data analysis options', illustrated with 'survival times after cardiac allografts', taken from a classic article on survival post heart-transplant. That section, and the associated computer code, can be found on the author's website http://www.medicine.mcgill.ca/epidemiology/hanley/software.

In that material we show how we would deal with these data today, using time-dependent covariates in a multivariable parametric or semi-parametric (hazard) regression model, with subjects switching 'exposure' categories (states) over time. However, we found it instructive to begin with the classical approaches already widespread in 1969, in particular those in Mantel's classic 1959 article. We use his 1974 generalization of lifetables[18] to deal with transitions between 'exposure' states; indeed, Mantel's 1974 paper is the conceptual forerunner of what is now known as regression for 'time-varying covariates'. We also estimate the mortality rates and rate ratios using Poisson regression.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada, and by the Canadian Institutes of Health Research.

Conflict of interest: None declared.

Postscript

When writing this piece, we wondered whether we were preaching to the converted. We did not add Rodolfo Saracci's suggested subtitle, 'The fallacy that refuses to die'. Surely such blunders do not occur in epidemiology journals, where the review is more rigorous than in some of the clinical ones? The article that is the subject of the correspondence in this IJE issue[19] indicates otherwise. The flaw in the comparison that led to a multifactorially adjusted, but too good to be true, hazard ratio of 0.52 (and even the other, more finely stratified ratios) was missed not just by the authors themselves, but also by their colleagues, granting agencies, journal referees and editors, and newspaper journalists and editors.

The Editor asked us to 'explain how immortal time bias plays a role in their findings' and to provide 'any comment [we] care to make about their re-analysis in response[20] to the criticisms raised by Lange and Keiding'.[21] We do so, but only after we first make some broader comments.

It will not be easy to put the toothpaste back in the tube, but we hope that those in the academic portion of this chain will each do their part. Might the *IJE* ask its media contacts to carry a follow-up story that might help undo the damage? In addition, instead of reporting additional analyses that still have flaws (or faulting the media for the over-interpretation and for their focus on the longevity 'effect') an *IJE mea (nostra?) culpa* might do more good: it might just add to (rather than subtract from) the limited amount of credibility biomedical scientists currently have remaining with the public.

It is one thing to give the public a reason to merely daydream about winning an Oscar and adding four years to one's life; it is quite another to imply—even cautiously—on the basis of the difference in median longevity of six years in the bottom left panel of Figure 1 of the 'sun exposure' article, that an even larger longevity bonus is readily accessible to all. Curiously, the 'extra' six years do not appear anywhere in the article, but figured prominently in the newspaper story. In it, one of the authors emphasized that they could not identify the direct causal link, but added that 'the numbers as such do not lie'. This statement illustrates what one might call a type III error, where an inappropriately set up statistical contrast, not chance, is the culprit.

We comment later on the less-emphasized, but possibly valid, results in Table 3 of the article, and first address the contrasts that led to the crude difference of six years and the 'adjusted' hazard ratio of 0.52. How could these analyses have received the most prominence, and without anyone ever raising an alarm? There was a hint that the authors understood, on some level, that such analyses involved immortal time: there is a statement about 'the temporality between the exposure and the outcomes'. The use of logistic regression for two of the outcomes, but Cox regression for the other, should also have prompted reviewers to try to understand why. In retrospect, the warning signs were all there: *P*-values so small that—even setting aside the concern already raised, tongue in cheek, about their numerical accuracy—they may be the smallest that have ever appeared in print anywhere; very different answers from the various data-analysis approaches; Kaplan–Meier curves and log-rank tests with no mention of staggered entry, but Cox regressions that do; the quite telling pattern of hazard ratios in the lower left panel of Figure 2; and, most importantly, at least to those not involved in the publication chain, the six years and even the adjusted hazard ratio of 0.52 seemed way too good to be true (in industrialized countries the mortality rate ratio (females:males) is approximately 0.7). On the other hand, as with the possibility of burnout of presidents who are early achievers, or of the extra health benefits of being rich, or of taking out a failed transplant, the 'more-(activity-in-the-)sun-is-good' hypothesis has a certain plausibility to it, and there is other evidence, based on the 10-year survival of those with basal cell carcinoma (Jensen *et al.*).[22] Moreover, the authors had used a clever (if somewhat unusual) way to study it, and used sophisticated statistical tools with extensive high quality data. The consistency in those below age 90 ('less MI, less fracture, less death') was taken as further evidence in support of the hypothesis.

One way to 'check' for immortal time bias is to study an event (outcome) that should have no causal relationship with the exposure of interest, and to be wary if the hazard ratios are clearly below 1. Alternatively, one may examine the association between a clearly 'unrelated' exposure and the outcome of interest, as we do below.

The article has conflicting descriptions of what was done to generate the less-emphasized results in Table 3. The statistical methods describe (synchronized?) matched sets, each comprising one 'exposed' person and five persons referred to as 'general population controls' but in the results section it is referred to as a 'matched case-control study'. Importantly, the authors do tell us that 'only myocardial infarction and hip fracture events following a diagnosis of non-melanoma skin cancer or cutaneous malignant melanoma entered into the analysis, whereas events before skin cancer were excluded'. This, together with the (adjusted) all-cause mortality hazard ratios of 0.96 and 0.97, suggest that, whatever they called it, their analysis may have—partially at least—avoided the 'temporality' problem (As we illustrate below, the crude six years, and the adjusted 0.52, and even the hazard ratios in the authors' additional analyses do not). The authors now realize that the analyses in Table 3, initially relegated to the very end of the Results and not discussed further, are probably the least biased. If one takes the fully adjusted 0.97 from the matched study as the closest to correct, one could put it into context for the newspaper readers by saying that it translates into a longevity difference of about four months rather than several years.

To show how immortal time bias plays a role in their findings, and to try to understand if the additional analyses are free of it, we examine the association between an unrelated exposure and death from any cause. Retroactively, and randomly, and without communicating the information to anyone, we choose a number of anonymous Danes in the Lexis rectangle enclosed by ages 40–110 years and calendar years 1980–2006 to be 'prizewinners'; winning the prize is the new, and obviously 'irrelevant' exposure. The population size Lexis dataset available in the Human Mortality Database (http://www.mortality.org) has a total of 4 130 227 Danes (the survivors, past age 40 and past 1980, of 91 different birth cohorts) in the leftmost column or bottom row of the rectangle. Using R code (available on our website) we simulated a yearly lottery that selected some of them to be virtual prizewinners. The incidence of prizewinners was an age-function with the same shape as the age-specific incidence of non-melanoma skin cancer in several Canadian provinces, scaled (downwards!) so that the total number of winners, and the average age of winning, were close to the 129 000 cases of skin-cancer, and the average diagnosis age of 68, in the *IJE* article. The only condition was that the winner had to be alive at the time of each yearly draw. Unlike other lotteries, there was, in large print, a statement that 'no other conditions apply'.

By its nature, the prize could not extend their longevity. Yet, just because of this 'must be living' condition, when we used the same analysis as in Figure 1 in the *IJE* article, we obtained a difference in median longevity of 8.5 years (and a hazard ratio of 0.57 with a *P*-value somewhere below the R pchisq function limit of 5×10^{-324}). Moreover, the hazard ratios we found in the 10-year 'strata' looked very similar to those in the lower left panel in the *IJE* Figure 2. Furthermore, when (as the authors do in their response) we narrowed the age slices further and insisted that 'those who [won our prize] beyond the age-strata enter into the analysis as not having [won]', we again get patterns similar to those in the figure in the

response to Lange and Keiding. Even using age-slices just two years wide, our hazard ratios were not null: they ranged from 0.93 at age 65 to 0.95 at age 85. The reason for the residual bias is that, by definition, a person who receives the prize at age 77.9 is 'immortal' for 1.9 years of the 2-year age slice 76–78. To avoid this induced immortality entirely, one needs to shrink the age-slice to an instant. Doing so is equivalent to using a time-dependent covariate ('exposure') in the Cox model, with risk sets defined at the moments the events occur. This is the most common way to deal with exposure states rather than traits.

By matching the Cox models on age, the authors did compare people who have survived to the same age. This may have led them and the reviewers to think that all was now taken care of. But with changing exposures, age-matching alone is not sufficient: one must also properly identify and update each subject's unexposure status as he/she proceeds through time and through the risk sets.

Sadly, we must add one more example to the list begun by Farr: to enter the index category, i.e. be promoted in one's profession; enter the list of cricket or jazz greats; enter the period of exposure to risk; receive an organ transplant; have it removed; be prescribed inhaled corticosteroids or a statin; win an Oscar; or receive a diagnosis of non-melanoma skin cancer, one needs to have lived long enough (in the reference category) in order to do so. No such minimum longevity requirement is imposed on entry to the reference category itself.

Annotated references

1. Ayus JC, Achinger SG, Lee S, Sayegh MH, Go AS. Transplant nephrectomy improves survival following a failed renal allograft. *J Am Soc Nephrol* 2010;**21**:374–80. See also the editorial in the same issue. We provide extensive commentary in sections 2 of our article, and in the Supplementary material.
2. Farr W. *Vital Statistics*. A memorial volume of selections from the reports and writings of William Farr. London: The Sanitary Institute, 1885.
   The now-easy access, clear writing, and opportunities to compare the concepts and principles with those in modern textbooks, make several portions of this volume worth studying even today.
3. Hill AB. Principles of medical statistics. XIV: Further fallacies and difficulties. *Lancet* 1937;**229**:825–27.
   'The average age at death is not often a particularly useful measure. Between one occupational group and another it may be grossly misleading … the average age at death in an occupation must, of course, depend in part upon the age of entry to that occupation and the age of exit from it—if exit takes place for other reasons than death'.
4. Hill AB. Principles of medical statistics, XII: Common fallacies and difficulties. *Lancet* 1937;**229**:706–08.
   With the dates changed, the worked example could easily pass for a modern one: 'Suppose on Jan. 1st, 1936, there are 5000 persons under observation, none of whom are inoculated; that 300 are inoculated on April 1st, a further 600 on July 1st, and another 100 on Oct. 1st. At the end of the year there are, therefore, 1000 inoculated persons and 4000 still uninoculated. During the year there were registered 110 attacks amongst the inoculated persons and 890 amongst the uninoculated, a result apparently very favourable to inoculation. This result, however, must be reached even if inoculation is completely valueless, for no account has been taken of the unequal lengths of time over which the two groups were exposed. None of the 1000 persons in the inoculated group were exposed to risk for the whole of the year but only for some fraction of it; for a proportion of the year they belong to the uninoculated group and must be counted in that group for an appropriate length of time.'
   A mathematical proof that 'neglect of the durations of exposure to risk must lead to fallacious results and must favour the inoculated' can be found in the paper by Beyersmann *et al.*, *J Clin Epidemiol* 2008;**61**:1216–21. Hill goes on to describe a 'cruder neglect of the time-factor [that] sometimes appears in print, and may be illustrated as follows. In 1930 a new form of treatment is introduced and applied to patients seen between 1930 and 1935. The proportion of patients still alive at the end of 1935 is calculated. This figure is compared with the proportion of patients still alive at the end of 1935 who were treated in 1925–29, prior to the introduction of the new treatment. Such a comparison is, of course, inadmissible'. Today's readers are encouraged to compare their reason why with that given by Hill.
5. Hill AB. Cricket and its relation to the duration of life. *Lancet* 1927;949–950.
6. Breslow NE, Day NE. *Statistical Methods in Cancer Research: Volume II: The Design and Analysis of Cohort Studies*. New York: Oxford University Press, 1994.
   The basis for Figure 3, and the source of the principle by which person-time should be allocated.
7. Duck BW, Carter JT, Coombes EJ. Mortality study of workers in a polyvinyl-chloride production plant. *Lancet* 1975;**2**:1197–99. 'Age-standardised mortality-rates for a population of 2100 male workers exposed to vinyl chloride for periods of up to 27 years do not show any excess of total or cause-specific mortality'.
8. Wagoner JK, Infante PF, Saracci R. Vinyl chloride and mortality? [Letter] *Lancet* 1976;**2**:194–95.
   This letter, a response to Duck *et al.,* is notable both for its tongue-in-cheek 'possible interpretations' of the SMRs of 112, 107 and 61 (!) in the PY where exposure had been for <10, 10–15 and 15+ years, respectively, and for its re-allocation of the PY giving new SMRs of 79, 137 and 353!
9. Walker AM. *Observation and Inference: An Introduction to the Methods of Epidemiology*. Chestnut Hill, MA: Epidemiology Resources, 1991.
   This is the first author we know of to define and put the term 'immortal time' in writing. Readers are invited to compare the definition with the description provided by Bradford Hill.
10. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd edn. Philadelphia, PA: Lippincott-Raven 1998.
11. Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. *Am J Respir Crit Care Med* 2003;**168**:49–53.

The first author to put the term 'immortal time' in the title of an article.

12. Shariff SZ, Cuerden, MS, Jain AK, Garg AX. The secret of immortal time bias in epidemiologic studies. *J Am Soc Nephrol* 2008;**19**:841–43.
    This article appeared one year before, in the same journal as the one on transplant nephrectomy.
13. McCann SJH. The precocity-longevity hypothesis: earlier peaks in career achievement predict shorter lives. *Pers Soc Psychol Bull* 2001;**27**:1429–39.
    See section 4, and see two subsequent articles by McCann. Does the fact that President Ronald Reagan lived to be 93 support this hypothesis? 'Likewise, Stephen Senn (personal communication) suggests that the 'higher mortality rates' in 'the childless' (Agerbo *et al*. Childlessness, parental mortality and psychiatric illness: a natural experiment based on in vitro fertility treatment and adoption. *J Epidemiol Community Health* 2012;**00**:1–3), could equally be reported under the headline 'Those who die young have fewer children'.
14. http://www.nytimes.com/2010/08/31/science/31profile.html see also http://test.causeweb.org/wiki/chance/index.php/Oscar_winners_do_not_live_longer
15. Matthew Herper. Does winning an Oscar make you live longer? Three important lessons from a raging scientific debate. *Forbes Magazine*. Feb 26, 2012.
16. Leibovici L. Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *BMJ* 2001;**323**:1450–51.
    We modelled our simulation on this (*BMJ* Christmas Edition) article, which generated many intense responses. It began with the statement 'As we cannot assume a priori that time is linear, as we perceive it, or that God is limited by a linear time, as we are, the intervention was carried out 4–10 years after the patients' infection and hospitalisation'.
17. Turnbull BW, Brown BW, Hu M. Survivorship analysis of heart-transplant data. *J Am Stat Assoc* 1974;**69**:74–80.
    Their permutation test motivated us to simulate, by lottery, a virtual, after-the-fact and ineffective intervention for those with a failed allograft. It is a useful technique for quantifying how much of the longevity advantage is an artefact. See our earlier use of this strategy in Sylvestre M-P, Huszti E, Hanley JA. Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. *Annals of Internal Medicine* 2006; 145:361–63.
18. Mantel N, Byar DP, Evaluation of response-time data involving transient states illustration using heart-transplant data. *Journal of the American Statistical Association* 1974;**69**:81–86.
19. Jacobsen P, Nordestgaard B, Nielsen S, Benn M. Skin cancer as a marker of sun exposure associates with myocardial infarction, hip fracture, and death from any cause. *Int J Epidemiol* 2013;**42**:1486–96.
20. Brøndum-Jacobsen P, Nordestgaard BG, Nielsen SF, Benn M. Authors' Response to: Skin cancer as a marker of sun exposure—a case of serious immortality bias. *Int J Epidemiol* 2014;**43**:972–73.
21. Lange T, Kielding N. Letter: Skin cancer as a marker of sun exposure: a case of serious immortality bias. *Int J Epidemiol* 2014;**43**:971.
22. Jensen AØ, Lamberg AL, Jacobsen JB, Braae Olesen A, Sørensen HT. Non-melanoma skin cancer and ten-year all-cause mortality: a population-based cohort study. *Acta Derm Venereol* 2010;**90**:362–67

Additional references

Redelmeier DA, Singh SM. Survival in Academy Award-winning actors and actresses. *Ann Intern Med* 2001;**134**:955–62.

The widely reported almost four-year longevity advantage over their 'nominated but never won' peers includes the immortal years between being nominated and winning. Use of the years between birth and nomination is an example of many researchers' reluctance to subdivide each person's relevant experience. See the re-analyses by Sylvestre *et al.* (2006) in the same journal, and by Wolkewitz *et al.* (*Am Statistician* 2010;**64**:205–11) and Han *et al.* (*Applied Statistics* 2011;**5**:746–72).

Mantel N, Byar DP. Evaluation of response-time data involving transient states—illustration using heart-transplant data. *J Am Stat Assoc* 1974;**69**:81–86.

In 1972, Gail had identified several biases in the first reports from Houston and Stanford. One was the fact that 'patients in the [transplanted] group are guaranteed (by definition) to have survived at least until a donor was available, and this grace period has been implicitly added into [their] survival time'. Mantel was one of the first to suggest statistical methodologies for avoiding what he termed this 'time-to-treatment' bias, where 'the survival of treated patients is compared with that of untreated controls, results from a failure to make allowance for the fact that the treated patients must have at least survived from time of diagnosis to time of treatment, while no such requirement obtained for their untreated controls'. He introduced the idea of crossing over from one life table ('waiting for a transplant' state) to another ('post-transplant') and make comparisons matched on day since entering the waitlist. Incidentally, Mantel's choice of the word 'guarantee' is not arbitrary: textbooks on survival data refer to a 'guarantee time' such that the event of interest many not occur until a threshold time is attained. In oncology trials, a common error—usually referred to as 'time-to-response' or 'guarantee-time' bias—is to attribute the longer survival of 'responders' than 'nonresponders' entirely to the therapy, and to ignore the fact that, by definition, responders have to live long enough for a response to be noted (see Anderson, *J Clin Oncol*. 1983;**1**:710–19, and, more recently, Giobbie-Hurder *et al*. *J Clin Oncol* 2013;**31**:2963–69).

Glesby MJ, Hoover DR. Survivor treatment selection bias in observational studies: examples from the AIDS literature. *Ann Intern Med* 1996;**124**:999–1005.

'Patients who live longer have more opportunities to select treatment; those who die earlier may be untreated by default' and their three words 'survivor treatment selection', to describe the bias explain why some person-time is 'immortal'.

Wolkewitz M, Allignol A, Harbarth S, de Angelis G, Schumacher M, Beyersmann J. Time-dependent study entries

and exposures in cohort studies can easily be sources of different and avoidable types of bias. *J Clin Epidemiol* 2012;**65**:1171–80.

Using an example from hospital epidemiology, the authors give 'innovative and easy-to-understand graphical presentations of how these biases corrupt the analyses via distorted time-at-risk'. See also: Schumacher *et al*. Hospital-acquired infections—appropriate statistical treatment is urgently needed! *Int J Epidemiol* 2013;**42**:1502–08.

Rothman KJ. Longevity of jazz musicians: flawed analysis. [Letter]. *Am J Public Health* 1992;**82**:761.

A letter in response to a retired professor of management, and jazz amateur (but sadly also a statistical amateur), whose data analysis suggested that jazz musicians, despite their rough lifestyle, live at least as long as their peers. In 'Premature death in jazz musicians: fact or fiction?' (Spencer FJ. *Am J Public Health* 1991;**81**:804–05), the longevity of their peers was measured by the life expectancy of those born the same year as they, although the musicians are, by definition, immortal until they became musicians and eminent enough to be included in the sample. The tone of the letter provides also an interesting contrast with Farr's teaching style. See: Bellis *et al*. Elvis to Eminem: quantifying the price of fame through early mortality of European and North American rock and pop stars. *J Epidemiol Community Health* 2007;**61**:896–901; Hanley *et al*. How long did their hearts go on? A Titanic study. *BMJ* 2003;**327**:1457; Abel *et al*. The longevity of Baseball Hall of Famers compared to other players. *Death Studies* 2005;**29**:959–63.); Redelmeier *et al*. Death rates of medical school class presidents. *Soc Sci Med* 2004;**58**:2537–43; and Olshansky SJ. Aging of US Presidents. *JAMA* 2011;**306**:2328–29, for more appropriate ways to carry out such longevity comparisons.

van Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol* 2004;**57**:672–82.

They gave immortal time bias a slightly different name because they covered a slightly broader spectrum of situations. Their review surveyed articles containing survival analysis that may have incorrectly handled what they define as a 'baseline immeasurable' time-dependent variable, i.e. one that could not be measured at baseline. They focused not just on the exposure of interest, but also other time-dependent covariates. They describe an interesting study on whether patients having a follow- up visit with a physician who had received the discharge summary would have a lower rate of re-hospitalization. When analysed as a fixed-in-time variable (i.e. as two 'groups', we found a large difference in readmission rates. However, this is a biased association, because patients who are readmitted to the hospital early after discharge do not have a chance to see such physicians and are placed in the 'no-summary' group. When a (correct) time-dependent analysis is used, we found a much smaller rate difference. They examined a large number of observational studies that used a survival analysis, including the one on the survival of Oscar nominees and winners. The only 'baseline immeasurable' time-dependent covariate in that study the reviewer(s)identified was whether actors changed their names after baseline; missed was the fact that some nominees who did not win the first time they were nominated changed to the winners category subsequently. So, they 'cleared' this article, declaring it to be free of any possible time-dependent bias. Interestingly, they also thanked one of the authors of the Oscar study for 'comments regarding previous versions of this study'.

Carrieri MP, Serraino D. Longevity of popes and artists between the 13th and the 19th century. *Int J Epidemiol* 2005;**34**:1435–36.

Compared with the bishops in Farr's example, popes must have survived even longer just to become pope. Even though the authors were aware that longevity is a 'necessary condition for being elected Pope', their statistical approach did not fully address this constraint. Ideally, for each papacy-specific 'longevity competition', the time clock starts when the pope is elected, and the competition should include the pope, and those artists born the same year as him, who were still alive when he was elected. However, for several papacies, such detailed matching is not possible. Instead, for each of the 1200–1599 papacies, their analysis effectively 'started the clock' at age 39—the age at which the youngest pope in that era was elected—by excluding artists who died before reaching that age. For the 1600–1900 papacies, it was started at age 38. A re-analysis (Hanley JA, Carrieri MP, Serraino D. Statistical fallibility and the longevity of popes: William Farr meets Wilhelm Lexis. *Int J Epidemiol* 2006;**35**:802–05), that used a papacy-specific time clock for each papacy-specific longevity competition, reversed the original findings.