Selection into the NHIS: is the statistical cure worse than the bias?

Supplementary material

James A. Hanley
McGill University
James.Hanley@McGill.ca

| WEB APPENDIX | TOPIC | PAGES |
|--------------|-------|-------|
| 1            | THE MIETTINEN VS. THE LEVIN VERSION OF THE PAF FORMULA | 2-7 |
| 2            | CONTEXTS WHERE PAFs ARE/ARE NOT RELEVANT | 8-9 |
| 3            | AGE-SPECIFIC HRs DERIVED FROM NHIS DATA | 10-17 |
| 4            | IN SIMULATIONS, HOW WELL DID THE APPROACH OF MASTERS ET AL. CORRECT HRs FOR DELIBERATELY INTRODUCED AGE-AND-HEALTH-RELATED NON-PARTICIPATION? | 18-23 |
| 5            | R CODE TO PRODUCE THE 2 SIMULATIONS | 24-34 |

| WEB FIGURE | TOPIC | PAGES |
|------------|-------|-------|
| 1          | RESULTS FROM JH’S RE-ANALYSIS OF SELECTED NHIS DATA | 11 |
| 2          | AGE-SPECIFIC PATTERNS IN MALE: FEMALE HRs IN 6 EUROPEAN COUNTRIES | 19 |
| 3          | ATTEMPT TO RECOVER CORRECT AGE-SPECIFIC PATTERNS IN MALE: FEMALE HRs, AFTER INCREASING-WITH-AGE SELECTIVITY WAS INTRODUCED | 20 |
WEB APPENDIX 1: THE MIETTINEN VS. THE LEVIN VERSION OF THE PAF FORMULA

Introductory remarks

Some of the material below is drawn from JECH2001\(^1\) – a didactic article prompted by what I perceived as a serious lack of understanding of the logic and of the correct/incorrect use of both the Miettinen\(^2\) and Levin\(^3\) formulae. This poor understanding, particularly when the marker of increased risk is polytomous, or when there is confounding, still persists. This is despite a very clear non-algebraic explanation -- on pp. 53-55 of the widely read 2002 introductory text\(^4\) -- of the Miettinen version for an all-or-none marker, and -- in 2001 -- a comprehensive review paper\(^5\) that focused on regression contexts.

If – by say adding a further subscript -- one were to make each \(pd_{ik}\) specific to a BMI category, the structure of equation 4 in the AJPH 2013 article by Masters et al. would resemble the Miettinen version. In subsequent correspondence concerning the displayed equation, they indicated that they had computed the PAFs using the Levin version.

The Levin version of the PAF takes the form of “excess”/“(background+excess”); the Miettinen version takes the form “fraction in which some excess is possible” × “actual excess within this fraction”. Although few would give the resulting fraction or percentage a causal interpretation, the arithmetic equivalency of these two ways to calculate the PAF in the all-or-none case, and the different approaches behind them, can be illustrated by the following exercise suitable for an elementary school lesson on fractions. In 2014 in the USA, there were in all, some 1.30 million deaths among women, and 1.33 million deaths among men. Some 28 thousand deaths were from prostate cancer. What fraction of all 2014 deaths in the USA was

---

1 JECH2001: Hanley JA. A heuristic approach to the formulas for population attributable fraction. *Journal of Epidemiology and Community Health*. 2001; 55:508-514.
2 Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology* 1974; 99(5): 325-332.
3 Levin ML. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* 1953;9:531-541.
4 Rothman KJ. Epidemiology: an introduction. Oxford University Press, New York. 2002.
5 Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Statistics in Medicine* 1991;10:1753-1773.
attributable to prostate cancer? In the one (more Levin-like) version, the answer is \( P/(W+M) \); in the other (more Miettinen-like) it is \( (M/[W+M]) \times (P/M) \). They yield the same answer, 0.011, or 1.1%.

**Notes on these two PAF versions**

*“It’s obvious”*

In 2001, once I too finally understood how simple Miettinen’s ‘fraction of a fraction’ version really was, I published (JECH2001) an extensive ‘translation’ of his ‘etiologic fractions’ (EF) article, aided by numbers and diagrams to allow the reader to see how readily one can go between the two versions. When I asked Miettinen why he did not explain his version more fully than he tried to do in the second paragraph of the Discussion of that 1974 paper of his,

> That the (natural) preoccupation with the frequency of the factor in the source population of cases may be replaced by its frequency among the cases themselves may not be immediately obvious. However, if among those in the source population who have the factor a fraction \( EF_1 \) of cases is attributable to the factor, then this fraction \( EF_1 \) of cases with the factor are attributable to the factor, while none of the cases without the factor are. This equivalence implies formulas 3 and 4.

he replied that it took him three agonizing days of algebra to get from Levin’s version to his. He then looked at its ‘fraction of a fraction’ form, and declared to himself “it’s obvious”. Even today, just as I noted at the end in JECH2001, the simplicity and ‘immediate obviousness’ of its structure continue to be under-appreciated.

More recently I asked Miettinen how he would explain it today. He suggested an example of a lawyer in a suit brought by lung cancer victims against his ‘big tobacco’ client. That lawyer would immediately ask that the fraction of victims who were never-smokers be removed from the class of all lung cancers – their ‘EF’ is zero (‘none of the cases without the factor are attributable to the factor’). In equation 11 in his 1974 article, the summation is over all categories, including the referent one, but ‘from the referent category the contribution is zero’). Once the ‘cannot be attributed’ cases are removed, the focus is on what fraction of the remaining victims had their cancer caused by smoking (the etiologic fraction within the remaining faction).

As I related in JECH2001, Miettinen was quite cryptic in his 1974 paper, and did not give an algebraic derivation. He did cite another source which in turn cited an unpublished source. He
derived the “fraction of a fraction” formula 11 years later (equation A.2.17, page 256 of his book (Miettinen, OS. Theoretical epidemiology: principles of occurrence research in medicine. New York: Wiley; 1985) but in a seemingly different context, and using a purely algebraic trick that does not reveal the logic behind it. The appendix to JECH2001 gives the algebra.

PAF components

As the Masters et al. AJPH article\(^6\) (AJPH2013) re-iterates, the PAF is an amalgam of two quantities. One of these is the hazard ratio (HR) or rate ratio, which we might -- if we were flexible -- consider as being a parameter of nature, and not so easily altered/modified by person, place or time. In reality, such settings are difficult to imagine, since most HRs are themselves modified somewhat – or maybe even quite substantially – by these factors.\(^7\)

It is much more difficult to treat the second input to the PAF – the prevalence of each of the risk factor categories, or the amount of population-time in each category – as a parameter of nature. Even if the HRs associated with the marker could be thought of as reasonably constant over different populations, times and places, it is unlikely that the prevalence/distribution of the factor would be constant over these varied settings. Thus, any calculated PAF is quite specific to a particular setting or domain.

Which version is more versatile?

Despite its clumsier computational form, the Levin formula is somewhat more versatile, since one can couple any postulated (or estimated) HR in one setting with any arbitrary marker in another (i.e., in what Flegal et al. in AJE in 2004\(^8\) refer to as a ‘target’ population). The simpler

---

\(^6\) AJPH2013. Masters RK, Reither EN, Powers DA, Yang YC, Burger AE, Link BG. The impact of obesity on US mortality levels: the importance of age and cohort factors in population estimates. *American Journal of Public Health*. 2013;103:1895–1901. doi:10.2105/AJPH.2013.301379.

\(^7\) In a ‘from personal-experience’ footnote to his book with Doll on the causes of cancer, Richard Peto has an interesting explanation of why the per-cigarette HR associated with lung cancer was higher in England than in the U.S. The English smoked their cigarettes down closer to the end than the Americans did: when they were young, Peto and his young pals who collected and recycled the tobacco in discarded butts looked forward to the visits of American sailors to his home town.

\(^8\) AJE2004. Flegal KM, Graubard BI, and Williamson DF. Methods of calculating deaths attributable to obesity. *American Journal of Epidemiology* 2004;160:331-338.
Miettinen case-based form is necessarily limited to the particular risk factor prevalence in the setting in which the HR was estimated (what Flegal et al. refer to the ‘derivation sample’).

Interestingly, the Levin formula does not work in a matched ‘case-control’ study, since the distribution of the marker in the matched control-series (denominator–series) does not necessarily represent the distribution of the marker in the base from which the cases arose. In contrast, it is in such matched settings that the Miettinen formula is at its most elegant and resourceful. Indeed, the impetus for Miettinen’s 3 days of algebra in 1974 was to arrive at a formula that could be computed from just such studies (for Miettinen, a case-control study is merely one where the person-time denominators for the marker categories are unknown and must therefore be estimated from a denominator series). But, as is made clear by the diagrams in JECH2001, and by the Benichou review, Miettinen’s case-based version also applies directly to known-population-time-denominators, such as in the follow-up of the NHIS. As is evident in JECH2001, all of the arguments go through, and as shown in Table 3 there, it is the stratum-specific arithmetic, not whether the ‘S’ (source-based) or ‘C’ (case-based) formula is used, that is the key to understanding how confounding is to be dealt with.

Miettinen’s formula was extended (indeed rediscovered) by Bruzzi in 1985.9 As Bruzzi, Benichou and I all have illustrated, the case-based version also works well in complex multivariable regression settings since it avoids the need to categorize quantitative risk markers – both the marker of interest, and factors that, if ignored, merely confound. Indeed, as is explained in the last paragraph of JECH2001, one can continue to dis-aggregate the numbers of cases until one is dealing with ‘singletons’ where each ‘case’ defines its own exposure category (this could happen if the exposure takes on values along a continuum). Curiously, this attractive case-based version was not considered in Flegal et al.’s AJE2004 article (cited by Masters et al.) on ‘methods for calculating [the numbers of] deaths attributable to obesity’. I wonder if this is because of the limitations of the Miettinen version and the greater flexibility and transferability of the Levin version.

9 Bruzzi P, Green SB, Byar DP. Estimating the population attributable risk for multiple factors using case control data. American Journal of Epidemiology 1985;122:904-914.
Which version did Masters et al. use?

Equation 4 in AJPH2013 looks like the Miettinen version, but it seems that, contrary to the impression given by the display of that version, Masters et al. actually used a Levin version in their calculations. This confusion, and my difficulty in following all of what they did do, is one of the reasons I suggest that editors and reviewers focus less on which version of SAS or Stata the authors used, and more on having the computer code itself available to reviewers and readers.

Effect modification

One item not reviewed in JECH2001, but partially covered in AJE2004, is the ‘effect modification’ context (Table 3 in JECH2001 assumed a common RR of 1.5). A general version-independent understanding of this more general context – one that is also covered in the Masters et al. exposé – follows quite naturally by modifying the HRs in Table 3 in JECH2001: simply change the highest rate ratio in that Table from 3 to 4 say, and confirm that the total excess number (now 85) is now 36% of the total number of cases (now 235). One can obtain this by either formula, provided one does the arithmetic separately for each stratum: the new EF₁ is 4/9ths rather than 2/7ths, and it applies to 180 rather than 140 cases; the Levin fractions of 1/11 and 4/9 are applied to the 55 and 180.

A principle that helps in the understanding of both computational versions

I deliberately began this appendix with an unusual numerical example. I did so in part to emphasize the simplicity of Miettinen’s algebra-free version. But I also did so to raise questions most epidemiologists instinctively do when confronted with numerator-only data: how large are the respective male and female USA populations? shouldn’t one use (incidence-density-type) rates?¹⁰ Even though the example is unusual, it shows an important aspect of PAFs. An overall

---

¹⁰ Depending on the type of data available, PAFs may be computed using ratios of person-based risks (proportions) or population-time-based rates (incidence densities, hazards). When dealing with a large portion of the lifespan, the distinction between risks (for all cause mortality, the lifetime risk is, as Mark Twain noted, always 1) and rates -- and thus between their ratios -- becomes important.

Incidentally, the population-time denominators underlying the 1.30 and 1.33 million deaths are 162 million woman-years and 157 million man-years respectively. Also, many of the ‘relatives’
death count, that in any other epidemiologic context would serve as a numerator input to a rate, instead serves as the (conceptual) PAF *denominator*; the sub-counts serve as the numerators and denominators for the sub-fractions used to compute the overall AF.

One comes to appreciate this ‘the denominator of any PAF is an overall number of cases’ principle better if one examines the PAF formulae used when there are confounding variables. No matter which version, one must disaggregate the numbers of *cases* (deaths) by covariate (confounder) patterns. One then computes either the case-based or exposure-based etiologic fraction for each separate pattern, before finally aggregating the sub-counts ‘attributable’ to the marker. This is particularly important with the more complex exposure-based version. And, if there is effect modification, a different HR is used for each modifier level.

In the next appendix I refer to those who have given very serious thought to what, if anything, an ‘obesity-related’ PAF would mean. Some of the criteria they invoke also arise here if one were to try to give a *causal* interpretation to the 1.1% calculated above. It is an easy task to *numerically* compute an overall mortality rate in the absence of fatal prostate cancer, or if the rate of fatal breast cancer was the same in women as in men. It is *quite another* to prescribe how those in the higher risk category might achieve these lower rates.

______________________________
in the research on obesity are ratios of hazards or incidence densities, not of long-term risks.
WEB APPENDIX 2: CONTEXTS WHERE PAFs ARE/ARE NOT RELEVANT

Contexts where PAFs are relevant

Population attributable fractions might be relevant when the time at increased risk is limited to when a factor is present. Overall numbers (or rates) of such events might involve: motor vehicle crashes in persons driving at different alcohol levels, using/not using the cell phone, following extended/normal work hours, without/with daylight driving lights; strokes while on/not on blood thinners; or venous thromboembolism postpartum/not. The arithmetic consists of dividing the numbers of events in each increased-risk state into ‘background’ and (as depicted by the red-colored rectangles in Web Figure 1B below) ‘in excess of this’, and expressing the overall excess number (or rate) as a fraction of the total number (or rate). Non-experimental data make it difficult to avoid confounding by extraneous factors, so “it would usually be pretentious to offer any specific estimate of the [etiologic] fraction attributable to a given marker of increased risk” (Miettinen 1974). And even if confounding were not an issue, there are other conceptual problems.11

Contexts where PAFs are not relevant

Neither fractions nor the HR inputs to them are causally interpretable for factors/markers such as obesity that, along with others, change over many life-decades and have/are associated with downstream consequences.12 13 14. It is impossible to be precise about the several domains (point of departure, interventions to reduce risk, timeframe, other risk factors, … ) in the question, “what if people were to change, or could have changed, or we were able to change their environment/lifestyle?” It is no longer the simple arithmetical question of “what if the lowest

11 Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. American Journal of Epidemiology (1988). 128:1185-1197.
12 Levine BJ. The other causality question: estimating attributable fractions for obesity as a cause of mortality. International Journal of Obesity (2008) 32, S4–S7
13 Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. International Journal of Obesity (London). 2008 Aug;32 Suppl 3:S8-14. doi: 10.1038/ijo.2008.82.
14 Lajous M, Banack HR, Kaufman JS, Hernán MA. Should patients with chronic disease be told to gain weight? The obesity paradox and selection bias. American Journal of Medicine. 2015 Apr;128(4):334-6. doi: 10.1016/j.amjmed.2014.10.043. Epub 2014 Nov 22.
event rate were applied to all marker categories?”. Is it really possible to quantify the expected population-level health dividend (over some specified future time horizon) from encouraging individuals in higher weight categories now to move to (or towards) the normal weight category? How will individuals do so, and how will they hold all other factors (including those ‘included in the model’, and possibly, omitted ones such as tobacco and alcohol use) constant over this (presumably specified) portion of their future lives? Can we quantify the excess that has already occurred, but would not have have ‘if only’, at some specified time in their past, individuals had moved to(wards) a lower risk weight category?

Beyond PAFs and HRs

Car crashes, strokes and VTEs can be averted, and the appropriate PAF formulae for these contexts have long ago been addressed in detail -- even if they have not always been properly understood or applied (cf. Web Appendix 1). Deaths cannot be entirely averted, but they can be postponed. In contexts where the causal basis for the public health advice is solid, numbers of years lost\textsuperscript{15} or of premature deaths might be better communicators than hazard ratios.

\textsuperscript{15} Walton W, Dajnak D, Beevers S, Williams W, Watkiss P, Hunt A. Understanding the Health Impacts of Air Pollution in London. King’s College London. Final Report. July 14, 2015
WEB APPENDIX 3: AGE-SPECIFIC HRs DERIVED FROM NHIS DATA

Introduction

Unfortunately, to understand the data-analysis steps that produced the reported age-specific HRs reported and shown in Figure 2 in the AJPH2013 article, one needs to also read the correspondence following the earlier AJE article (AJE201316). In their more parsimonious AJPH2013 approach Masters et al. fitted smooth-in-age and smooth-in-time HR models: their AJE2013 work used several indicator (‘dummy’) variables to represent each one. But the smoothness is not the central issue. The central validity ‘contest’ is between the AJPH2013 equivalent of ‘AJE2013 Model 2’ (which does not try to ‘correct’ the HRs for a postulated selection bias) and ‘AJE2013 model 3’ or its smooth AJPH2013 counterpart (both of which do). If we take ‘truth’ to mean the HRs that would be seen in the ideal situation where everyone eligible to take part in the NHIS (or indeed everyone alive in the U.S.A. at the time) agreed to participate17, which is closer to this truth?

In going from AJE2013 to AJPH2013, Masters et al. also switched from Cox models to less-familiar Royston-Parmar survival models. Web Figure 1 below shows some re-analyses of the basic data from one illustrative demographic subset (white males) of the NHIS. I re-analyzed them to be sure that I (and other readers) would understood how Masters et al. ‘corrected for’ (what they believed was a substantial) increasing-with-age-and-ill-health rate of exclusion from or non-participation in the NHIS. I also wished to verify that the large ‘swing’ in the age-specific HRs (from attenuation-with-age HRs if one did not ‘correct’ to increasing-with-age HRs if one did) was not peculiar to a particular hazard model (Cox vs. Royston-Parmar vs. Poisson), but was specifically due to the use of age-at-entry to ‘remove’ the effects of the selection.

16 AJE2013. Masters RK, Powers DA, Link BG. Obesity and US mortality risk over the adult life course. American Journal of Epidemiology. 2013 Mar 1;177(5):431-42. doi: 10.1093/aje/kws325. Epub 2013 Feb 3.

17 I ignore errors in participants’ self-reported height and weight, and the ‘only at entry’ nature of BMI and the other relevant covariates. Also, the limited follow-up of some waves, and much longer follow-up of others, make it unclear what timescale a reported HR of 1.5 refers to: does it refer to short-term (1-year ahead) life-insurance premiums? or to annuities with a 20-year horizon?
A: Schematic of NHIS waves, and follow-up.
B: Numbers of deaths, hazard ratios, and implied etiologic fractions (red area as fraction of total area) for each of 4 BMI categories (shown from left to right).
C: Black curves (with line thickness indicating BMI category, and selected ages-at-entry) represent the fitted model (see equation [*] below). Red curves were calculated by then setting the (negative) coefficient associated with the product involving age-at-entry to zero, and retaining the (positive) coefficient associated with the product involving attained age. For more details, see text.
The ‘raw’ data

The purpose of Web Figure 1A is largely schematic, but it does use dots whose sizes (areas) are proportional to numbers of deaths. These are the numbers of deaths of white male NHIS participants at each attained age in each calendar year (the follow-up in the version of the NHIS data I used had been updated to 2011\textsuperscript{18}; Masters et al. has follow-up only to 2006), color-coded by decennial birth cohort. The parallelograms contain the ages attained by those from 5 selected surveys.

The broad patterns in the data are apparent in Web Figure 1B, which shows the numbers of deaths occurring in the person-years lived within the decade of attained age by NHIS participants who – at entry – reported being in the indicated weight category. The division of each x-axis is proportional to the numbers of person-years of follow-up in the categories and the areas of the rectangles are proportional to the numbers of deaths. The heights of the rectangles for the overweight and obese categories, relative to the height of the normal-weight rectangle, are Mantel-Haenszel hazard ratios, matched on one year of age and one year of the calendar\textsuperscript{19}.

If one wished to, the ‘excess’ numbers of deaths, shown as red rectangles, could be quickly calculated by the Miettinen version (see Web Appendix 1, or equation 4 in AJPH2013) by applying the two (HR-1)/HR fractions to the numbers of deaths in the two obese categories (in each age-decade, the death rates in the ‘overweight’ category were slightly lower than those in the ‘normal’ category). Clearly, if an overall fraction were to be calculated from them, it would be small. The reasons for this small fraction are three-fold: most deaths occurred in, and most of the person-years of follow-up were contributed by, participants who – at the outset – had reported being in the normal or overweight categories; the HRs in the overweight category are all below 1; the HRs in the two obese categories are all below 2 -- and in the age decades where most deaths occur, below 1.6 -- so that the (HR-1)/HR fractions (the ‘excess fractions’), applied

\textsuperscript{18} Minnesota Population Center and State Health Access Data Assistance Center, Integrated Health Interview Series: Version 6.11. Minneapolis: University of Minnesota, 2015. http://www.ihis.us

\textsuperscript{19} Since it was not clear to me which of the many other factors recorded in the NHIS should be ‘adjusted for’, I did not match on any of them.
to the numbers of deaths in these categories range from 0% to 49%, with the larger fractions applied to the smaller numbers of deaths.

If, as I did, one repeated the above analysis five times, each time restricting it to those born in a different decade, one would find that the excess numbers of deaths as a percentage of all deaths remain small.

I also fitted a simple Gompertz-Poisson regression model\(^\text{20-21}\) (with mortality rates log-linear in attained age, birth year, weight-category indicators, and attained age × weight-category-indicator products, so that the HRs are \textit{modified-only-by-attained-age}). Even though it was fitted with far fewer terms than their AJE2013 model 2, the HRs for the highest-risk weight category have the same ‘\textit{lower in older attained ages}’ pattern that one sees in the weight-category-specific HRs in Web Figure 1B, and that Masters et al. found with model 2 in their AJE2013 article: they were attenuated from approximately 1.8 at aged 50 to 1.4 at age 85, or somewhat less than 1% per year of attained age. So, with attained-age centered at 60, the equation was \textit{approximately fitted HR = exp[ 0.57 - 0.01 \times attained-age]}

\textbf{‘Corrected for selections bias’ age-specific HRs fitted using the Masters et al. approach}

Both model 3 in AJE2013, and the HR model in AJPH2013 were intended to correct for a postulated substantial (increasing with calendar time and age-at-entry) non-participation by selected subjects who were both less healthy, and more likely to be obese. To do so, Masters et al. included ‘age at survey’ terms in these statistical models. In my re-analysis, I did likewise, adding age-at entry, its square, and its product with each weigh-category indicator, to my simpler model. The coefficients in this model I fitted were very similar to the coefficients in theirs. Thus, in the highest weight category for example, with attained-age centered at 60, and age-at-entry at 48, I obtained the following function of attained age and age-at-entry

\[ \text{fitted HR = exp[ 0.57 - 0.01 \times attained-age]} \]

---

20 Fully parametric log-linear in age hazard models, such as these fitted by Benjamin Gompertz (“On the nature of the function expressive of the law of human mortality, and on a new mode of determining life contingencies” \textit{Phil. Trans. R. Soc. Lond.} 1825: 115, 513–583), still fit modern data reasonably well and still show the same approximately 8% increase in the force of mortality per year of attained age (HR=1.08) that he found almost 200 years ago, indicating that 1-year life insurance premiums should be about 8% higher at successive attained ages.

21 Breslow NE, Day NE. Statistical Methods in Cancer Research. Volume II - The Design and Analysis of Cohort Studies. IARC Scientific Publications No. 82. Chapter 4. Fitting Models to Grouped Data.
[*] fitted HR = exp \[ 0.57 + 0.02 \times \text{attained-age} - 0.03 \times \text{age-at-entry} \].

It is plotted as black segments in Web Figure1C. The pattern agrees with that seen in Web Figure1B: in those recruited when young, the obese:normal HRs were quite high throughout their years of follow-up [follow-up in the NHIS data I had available to me ranged from (2011-2004=) 7 to (2011-1986=) 25; 18 years shown in Web Figure1C]. In those recruited when older, the obese:normal HRs were much lower. Importantly, Web Figure1C reminds us that the range of ages attained during follow-up of any age-at-entry subgroup is necessarily limited (I used a few more years of follow-up data than Masters et al. did: in their data, the follow-up was 20 years from entry for wave 1, and only 2 for wave 19, so the mean range in their data is only about 10 years. Even in the additional data I had, one should regard a portion of each 18 year-follow-up segment shown in Web Figure1C as an extrapolation.

Up to this point, this fitted model is just that – it continues to fit (summarize) the patterns in the actual NHIS data. Even with the slight refinement (curvature) introduced within each segment, the broad age-related patterns of the HRs are still the same: for example, the HRs for the highest-risk weight category continue to be attenuated by somewhat less than 1% per year of attained age, whereas those for the second-highest weight category continue to be elevated but almost independent of attained age.

It is what Masters et al. did next, i.e. how they used the fitted model to ‘correct for’ selection bias, that created the very different age-pattern in their reported HRs. And it this now-increasing-with-attained-age pattern (now with large HRs in the age range where the majority of deaths occur) that altered their PAFs from a few percentage points to ones ranging from 10-20%. The large PAF differences were not due to which version of the PAF formula they used (see

---

22 If we were to take just the first year of follow-up, so that attained age was perfectly correlated with age-at-entry, the +0.02 \times \text{attained-age} and the -0.03 \times \text{attained-age} in this more complex equation would reduce it to the -0.01 \times \text{attained-age} in the simpler equation in the previous page.

23 Masters et al. did not respond to the letter from Wang Z and Liu M (Obesity–mortality association with age: wrong conclusion based on calculation error. AJPH July 2014, Vol 104, No. 7. Letters. pages e3-e4). Contrary to the impression that the non-response may have created, the HRs do not appear to be the result of a calculation error. Rather, as Masters et al. did explain in a response to a letter to AJE (Masters RK, Powers DA, Link BG. The authors reply. American Journal of Epidemiology 2014;179(4):529–532), they are the result of a deliberate attempt to ‘adjust for’ selection bias.
Web Appendix 1, or to their switch from the age-at-death distributions in the NHIS to those in the U.S.A. at large.\(^{24}\)

I too did what Masters et al. did: I took equation [*] above, and ‘set’ all of the ages at entry to the centered value of 48 years. This, I effectively suppressed the “age – 0.03 × age-at-entry” term, and obtained

\[
\text{‘corrected for selection bias’ HR} = \exp[ 0.57 + 0.02 \times \text{attained-age}].
\]

I then plotted this HR-curve (and the corresponding curves for the other two weight categories) in red in Web Figure 1C. They are quite similar to those in Figure 2 of AJPH2013.

**Comments**

As others have pointed out, just on statistical grounds alone, this fitting of equation [*] ‘sails very close to the wind’. In the fitted model there was a high correlation of almost –0.9 between the fitted coefficient for the products of the obesity-category-indicator with (1) attained age and with (2) age at entry. This was to be expected in light of the limited follow-up of each NHIS participant, making attained ages-at-follow a strong function of age-at-entry. Most readers faced with this degree of collinearity would find it impossible to say which of the two is the more genuine modifier of the HRs, and would not trust the large degree of statistical extrapolation. In this case, however, there are so many deaths that the confidence ellipse centered on the fitted pair of coefficients (+0.02, -0.03) is quite tight.

But, again, it is important to not get caught up on the purely statistical issues of collinearity. One needs to ask the more basic conceptual question: exactly what does such a ‘corrected for selection bias’ HR of \( \exp[ 0.57 + 0.02 \times (80-60)] = 2.6 \) at age 80 mean? The following questions suggest that its meaning may be quite elusive.

Imagine an ideal case where the participation rate in every wave was 100\%, and that in 1995 (the middlemost NHIS wave), regardless of their health, all of the white males aged 48 (the

\(^{24}\) In an all-or-none risk marker situation, the Miettinen formula (Appendix 3) makes it particularly easy to see how one can/cannot arrive at a PAF of 15\%. One can if some 30\% of all deaths occurred in persons in whom the marker is/was present [I prefer this wording over “the prevalence of obesity among the deceased”], and the HR is 2 (EF\(_1\) = 1/2 = 50\%). One can if some 20\% of all deaths occurred in persons who are/were in this category, and the HR is 4 (EF\(_1\) = 3/4 = 75\%). One cannot if only 10\% of all deaths occurred in persons in this category, no matter how elevated the HR is.
middlemost NHIS wave) who were eligible for that wave were categorized according to their BMI at that time, i.e., suppose that the mortality rates \( (m)'s \) in actuarial notation) for each of the 32 successive ages/years, up until age 80 in 2027, were calculated for each (1995) weight-category. Can we take the HR=2.6 to mean that the incidence density or force of mortality \( m_{80} \) in the index category at attained age 80\(^{25}\) would be 2.6 times higher than the corresponding one in the reference category?

The \( m_{80} \) for each 1995 category must be an amalgam of the \( m_{80-P} \) among those in the category who participated and the \( m_{80-NP} \) among who did not participate: does the fitted model tell us how big the ‘P’ person time is relative to the ‘NP’ person time, and how different the \( m_{80-P} \) and \( m_{80-NP} \) are? These numbers could provide a reality check and help epidemiologists explain the correction to a journalist.

In case the above interpretation is too restrictive, should the corrected HR = 2.6 at age 80 refer to a contrast that is still based on (corrected to) 100% participation, but on BMI categories established at entry at some (unspecified) past age or past calendar year?

How should insurance companies interpret the ‘corrected’ HRs for age 40? 60? 80?

Would it be better to use datasets such as from a (Framingham) cohort that recruited participants while still young -- before many of their birth cohorts had become unwell or died --, measured them regularly (so that the exact time horizons in the HRs are clear), and followed them for up to 67 years?

How close is this ‘correction for selection bias’ problem in the NHIS to the statistical challenges faced by Robins and others who have studied occupational health hazards,

\[25\text{ Although the time issues in this obesity context are different than those in a cross-sectional study, they remind me the advice, on p135 of J. Avorn’s book (Powerful medicines: the benefits, risks, and costs of prescription drugs. 2004, 1st ed, New York : Knopf) to be wary of inferences from one-time data: “It’s been observed that if you were to conduct a cross-sectional study of aging and ethnicity in south Florida, you might conclude that people there are born Hispanic, and then become Jewish as they get older.” Likewise, AF Beringause (on p38 of his 2001 book Aging: An encyclopedia for adding years to your life and life to your years) relates that “Leonard Hayflick likes to tell” that the conclusion would be that “Miami residents are born Hispanic and die Jewish.” I would add a similar time-related warning concerning a projection, from age 48 to age 80, based on a decade or so of actual follow-up, with no check along the way as to whether subjects had changed anything in their profiles.} \]
particularly in relation to ‘selection-in’ forces? What additional information and assumptions did they need to properly account for those who were too ill to enter employment? Would similar ‘causal inference’ tools be useful in investigations into the role of obesity in the etiology of premature mortality?

Even if the HR interpretation were not complicated by the causality issues already cited in Web Appendix 2, ‘adjusting’ the HR estimate for selection bias would already be a statistically challenging. Correcting a single HR for an imbalance in a known confounder in a sample where all of the scientifically selected subjects were measured, is one thing. It is quite another, as Masters et al. did, to use only a proxy (age-at-entry), and no other information, to correct a set of HRs (attained-age-specific) for the fact that some (unreported) fractions of the selected sample, for various unknown reasons, some presumably obesity- and health-related, some not, did not participate, and did not provide any data. The results of the simulations described in Web Appendix 4 below suggest that an adjustment/correction for ‘selection’ is much more elusive than that for a known confounder.

---

26 Provided, of course, the correlation between the confounder and the exposure of interest is not prohibitively high.

27 The weights in the NHIS might give some indication of how substantial, and how strongly age-related, the non-participation was.
WEB APPENDIX 4:  
IN SIMULATIONS, HOW WELL DID THE APPROACH OF MASTERS ET AL. CORRECT HRs FOR DELIBERATELY INTRODUCED AGE-AND-HEALTH-RELATED NON-PARTICIPATION?

Here I introduce selection bias into samples from known populations. One has ‘higher at older ages’ all-cause-mortality HRs, while the other has ‘lower at older ages’ ratios.

In the first of these, already shown in Journal Figure 1, which is based on an purely mathematical population, the “true” exposed : not exposed HR was approximately 1.1 at the lower and 2.5 at the upper age end. From it I simulated 19 sample waves where, increasingly with age at potential selection, exposed persons who did participate were less likely to die in the next six years than their sampled exposed peers who did not participate, thereby attenuating the observed HR curve. (The R code is given in Web Appendix 5.) Applied to these ‘selective’ data, the proposed Masters et al approach was not able to recover the HR pattern present in the unselected data. The ‘corrected’ age-specific HRs overshot the target at older ages, and undershot it at lower ages.
Attenuation of hazard ratios (HRs) with age has been seen in many cohort studies, such as in the age-specific smoker versus non-smoker contrasts in the British Doctors’ study, where 30% were over 50, and the overall participation was less than 70%. Thus some of that attenuation could be due to the selectivity considered here. But ‘lower at older ages’ all-cause-mortality HRs are also observed in contrasts in unselected populations: for example, as is seen in Web Figure 2 below, the male:female HR is typically > 2 at age 60; 1.5 at age 80; and near 1 at age 100.

Web Figure 2

---

28 Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 2016.01.02).
The curve in black in Web Figure 3 below shows the age-specific male:female HRs based the male and female cohort death rates in Finland. They are derived from the entire population.

**Web Figure 3**

| Birth Cohort | Web Fig. 3: Finland |
|--------------|---------------------|
| 1905         | 1915                |
| 1925         | 1935                |
| 1945         |                     |

Thus, in a second simulation, I used these, and the R code included in Web Appendix 6, to simulate 19 sample waves (1986-2004) from the entire Finnish population.

Whereas the only criterion for eligibility to be sampled was that persons be alive, I did vary the participation rates in these simulated samples: increasingly with age at selection, males who did participate in my samples were less likely to die in the next five years than their sampled male peers who did not participate (see Panel B of Figure 1 in the print version). The red dots in Web Figure 3 shows the age-specific male:female HRs computed from the data in these selective samples. Because of the selection-pattern I introduced, these HRs are quite a bit

---

29 Walker M, Shaper AG, Cook DG Non-participation and mortality in a prospective study of cardiovascular disease. *Journal of Epidemiology and Community Health*, 1987, 41, 295-299
lower than those in the unselected (‘population’) ones, particularly at older ages: while the male:female HR at age 40 is biased downwards by about 15%, the one at age 80 is biased downwards by about 50%.

The hope was that by using the age-at-entry recorded for each participant in these simulated waves, the Masters et al approach would be able to ‘recover’ the HR pattern in the entire (unselected) population, i.e., that it would be able to ‘adjust’/’move’ the red dots upwards until they were on top of the black curve.

However, when I applied this approach to the data from these ‘selective’ samples, I was not able to recover the true ‘lower at older ages’ HR pattern present in the unselected (i.e., the population) data. Instead, I obtained a curve, shown in purple in Web Figure 3, which ‘overshot’ its target at one end of the attained age scale, and ‘undershot’ it at the other. This ‘pivoting’ in the purple curve is similar to what one sees when one contrasts Figure 2 in Masters et al (AJPH2013, with ‘correction’ for selection) with their fitted model 2 in AJE2013 (no ‘correction’ for selection). As the interested reader may do using the R code provided, I was able to vary the degree of selectivity. The greater the selectivity (selection bias) I introduced, the more the ‘corrected’ curve of age-specific male:female HRs overshot its target at the upper ages, and undershot it the lower end – to the point that the ‘corrected’ male:female HRs at younger ages would be below 1. This would imply that younger men should demand that life insurance companies charge them \textit{lower} premiums that they charge younger women, and that older women should ask that the male:female premium ratios should be set much higher than they are.

\textbf{Concluding remarks}

The results of the simulations have reinforced my initial opinion – one widely held in epidemiology -- that trying to correct for selection bias is much more difficult than correcting for confounding. With a scientifically selected sample, a known confounder, no unknown confounders, and 100% participation, adjustment for confounding is a \textit{purely internal} matter of (synthetically) ‘redistributing’ all of these participants – or as Galton\textsuperscript{30} put it, leaving them in

\textsuperscript{30} Hanley JA. “Transmuting” women into men: Galton's family data on human stature. \textit{The American Statistician}, 58(3) 237-243. August 2004.
their original category and mathematically ‘transmuting’ their results (outcomes) -- so as to obtain a ‘fair’ contrast.

Since non-participation usually involves unseen elements, any resulting selection bias is usually thought of as an insidious aspect that cannot be overcome. A substantial portion of those selected decline to participate, for reasons suspected to be related to both the ‘exposure’ and the ‘outcome.’ Even if one knew the sizes of these missing portions, one would also need to know just how much sicker the non-participants were than the participants, and how these non-participants are distributed across the categories involved in the contrast. Using the ages-at-entry and the mortality pattern only in those who did participate to statistically infer/impute the relevant information on those who did not, appears to be asking the impossible from the available data.

As I have attempted to do, one first needs to test the correction method in a simple theoretical setting with a known target, so that one can assess by how much the proposed method might undershoot or overshoot it.

Some might be willing to overlook the causality objections raised by the theoreticians as to the estimability/measurability of the consequences of obesity. Wishing to not make the best the enemy of the good, they might still consider it meaningful/useful to calculate ‘bias-corrected’ HRs for BMI categories derived from once-only, self-reported height and weight data from the NHIS. And they might even dismiss the embarrassing corrections produced when the proposed bias-correction approach was applied to simulated samples.

Many more might consider it plausible that there was indeed some degree of increasing-with-age-at-entry, ill-health, and obesity-related non-participation in the NHIS, and thus that some degree of upwards-adjustment (statistical cure) is needed for the HRs seen at the older ages. But, what are they to make of the adjusted HRs at the younger ages shown in Figure 2 of AJPH2013, which are downwards-adjusted so much that they are well below 1? If these corrected HRs at these younger ages are correct, then younger people in the higher BMI

31 The HRs at these older ages -- where most people die -- dominate the PAFs calculations much more than which computational version of the PAF is used (see Web Appendix 1).
32 Few die at these ages, but many buy life insurance.
categories should ask life insurance companies to charge them considerably *lower* premiums than they charge those in the normal weight category.
R Code to Produce the 2 Simulations

```r
# Figure 1 for printed commentary, Jan 18 2017
# james.hanley@mcgill.ca

# entirely theoretical data
# using 8% rate inc. in force of mortality per year of age,
# and 'increasing with attained-age' mortality ratio (HR, male = 'exposed')
# HR = 1.1 or so at age 45 and 2.5 at age 85

base.age = 44
log.m.base.age = log(0.003)
base.year = 1985
base.gender = 1 # female ['exposed' = Male]

# force of mortality function
m = function(a, y, g) {
  log.h = log.m.base.age + 0.10 * (g - base.gender) +
         0.08 * (a - base.age) +
         0.02 * (a - base.age) * (g - base.gender) -
         0.01 * (y - base.year)
  return(exp(log.h))
}

# Deaths, PY, Age at Entry and Year at Entry
D = array(0, c(19, 110, 26, 2)) # wave (19..86-04) fu.year (26.. 84-11) age gender(1=ref, female)
PY = array(0, c(19, 110, 26, 2)) # PY
A.E = array(0, c(19, 110, 26, 2)) # Age at entry
Y.E = array(0, c(19, 110, 26, 2)) # Year at entry

entry.ages = 25:84 # 60
N.each = 100 # this many of each age at each wave
# not all that relevant, since use expected no.s of deaths

for(wave.year in 1986:2004) { # jan 1 each year,
  fu.years = wave.year:2011 # ends dec 31
  ny = length(fu.years)
  for(age.entry in entry.ages) {
    integral = c(0, 0)
    for(i in 1:ny) {
      fu.age = age.entry + i - 0.5 # mid y
      fu.year = wave.year + i - 0.5 # mid
      fu.index = i + (26 - ny)
      for(g in 1:2) {
        force.m = m(fu.age, fu.year, g)
        integral[g] = integral[g] + force.m
        PY[wave.year-base.year, fu.age, fu.index, g] = N.each * exp(-integral[g])
        D[wave.year-base.year, fu.age, fu.index, g] = force.m *
        PY[wave.year-base.year, fu.age, fu.index, g]
        A.E[wave.year-base.year, fu.age, fu.index, g] = age.entry
        Y.E[wave.year-base.year, fu.age, fu.index, g] = wave.year
      }
    }
  }
}

# Check: sum(D) ; sum(PY)
# check D[1:2,25:105,,1] ; PY[1:2,25:105,,1] ; A.E[1:2,25:105,,1] ; Y.E[1:2,25:105,,1]
# as vectors rather than arrays (for data analysis)
```
L=19*60+26*2  # 19 waves; 60 ages 40-99; 26 fu years, 2 exposure levels (ref, index)
AE=rep(NA,L)
YE=rep(NA,L)
A=rep(NA,L)
Y=rep(NA,L)
Male=rep(NA,L)

d=rep(0,L)
py=rep(0,L)

A.c = base.age

I=0
for(w in 1:19) {
  for(a in 40:99) {
    for(y in 1:26) {
      for(g in 1:2) {
        I=I+1
        A[I] = a - A.c
        Y[I] = y
        Male[I] = g-1
        py[I] = PY[w,a,y,g]
        d[I] = D[w,a,y,g]
        AE[I] = A.E[w,a,y,g]
        YE[I] = Y.E[w,a,y,g]
      }
    }
  }
}

ds = data.frame(A,Y,AE,YE,Male,py,d)
ds = ds[ds$py > 0,] ; # remove irrelevant cells
head(ds)

N = length(ds[,1]) ; N

ds$A.M = ds$A * ds$Male  # age*exposure product term, for effect modification by age

ds$o = log(ds$py) # for poisson regression

# check that regression recovers data-generation parameters
fit.unselected = glm(d ~ A + Y + Male + ds$A.M + offset(o), data=ds, family=poisson)

Ages=A.c:84

BETA = fit.unselected$coefficients[4:5]
HR = exp (BETA[1] + BETA[2]*(Ages-A.c) )

par( mfcol=c(1,1), mar = c(0.01,0.01,0.01,0.01), bty="n")

XLIM=c(34.5,89.5) ; YLIM=c(-3.65,4.15)

CEX=1.15

plot(Ages, HR , ylim=YLIM, type="l", col="black",
     xlab = "Attained Ages", yaxt="n", xaxt="n",
     ylab="Hazard Ratio (HR)", lwd=2.5, xlim=XLIM)

xx=40 ; segments(xx,0,xx,4.05) ; segments(xx,0,85.5,0)
for(ratio in seq(0,4,0.5)){
  segments(xx, ratio, xx-0.5,ratio)
text(xx-0.75,ratio,format(ratio,nsmall=1),
       cex=CEX,adj=c(1,0.5))
}
for(xa in seq(40,85,5)) {
  segments(xa, 0, xa, -0.05)
  text(xa,-0.1,toString(xa),adj=c(0.5,1),cex=CEX)
} text(xx-4.2,2,"Hazard Ratio",adj=c(0.5,0.5),srt=90,cex=CEX)
text((40+85.5)/2,-0.55,"Attained Age, years", adj=c(0.5,1),cex=CEX)
text(0.25+ Ages[length(Ages)], HR[length(Ages)], "Unselected", adj=c(0,0.5),col="black",cex=CEX)
segments(40,1,85.25,1,lty="dotted",lwd=2,col="grey20")
y=4.0
dy=0.3
x=41
FORAppendix=FALSE
if(FORAppendix) text(x,y,"Figure 2", adj=c(0,1))

COL="black"
rect(x-0.5,y-0.25*dy, 71, y - 6*dy,col="grey95",border=NA)
text(x,y-0.75*dy, "log of the Force of Mortality in unselected persons aged a, in year y, with all-or-none exposure e", adj=c(0,1),col=COL,cex=CEX,,family="mono")
text(x,y-4.15*dy, "= log[0.003] + 0.08*(a-44) + 0.10*e + 0.02*((a-44)*e) - 0.01*(y-1985)", adj=c(0,1),cex=CEX,family="mono")
#text(x,y-3*dy,"* Intake: starting 1986, 19 waves of 100 per #year-of-age, follow-up to 2011", adj=c(0,1),col=COL)

COL="black"
#text(x,y-4*dy,"* Selection: excludes some exposed persons who #will die within 6 years", adj=c(0,1),col=COL)

# selectivity
next.no.years = 6   # fn. to remove some deaths in males in these first # years after entry
fx.remaining = .85 # at entry.age 60 year.entry 1990 this fraction of deaths in Y of E retained
    # $$$ change to .99999 to remove almost all selectivity
b.0          = log(fx.remaining/(1-fx.remaining)) ;
b.age.entry  = -0.2   # fraction of deaths retained in these first years decreases with age at E # ie greater selectivity of healthier males at older ages
    # $$$ change to -0.00001 to remove almost all selectivity
b.year.entry = -0.02  # fraction of deaths retained decreases with year at E # ie greater selectivity of males at older ages
    # $$$ change to -0.00002 to remove almost all selectivity
fx.of.future.deaths.included = function(entry.age,entry.year,age) {
  LP = b.0 + b.age.entry*(entry.age-60) + b.year.entry*(entry.year-1990)
  F.0 = exp(LP)/(1+exp(LP))
  return( F.0 + ( (age-entry.age)/next.no.years ) * (1-F.0) )
}
# remove some deaths in male in first # years after entry
ds$fewer.d = ds$d   # to start with, then...
for (i in 1:N) {
years.ahead = ds$A[i] - (ds$AE[i]-A.c)
if( years.ahead < next.no.years & ds$Male[i]==1  ) {
    ds$fewer.d[i] = ds$fewer.d[i] * 
    fx.of.future.deaths.included(ds$AE[i], ds$YE[i], ds$A[i]+A.c)
}
}

head(ds)
sum(ds$d) ; sum(ds$fewer.d)  # overall numbers of deaths & py
sum(ds$py) ;

D = aggregate(ds$fewer.d, by = list(a=ds$A, m= ds$Male), sum)
Y = aggregate(ds$py, by = list(a=ds$A, m= ds$Male), sum)
H = D$x / Y$x
HR.empirical = H[61:120] / H[1:60]
age = D$a[1:60] + A.c
lines(ages[5:45],HR.empirical[5:45],col="red",cex=0.6,lwd=1.5)
    text(0.25+ages[45],HR.empirical[45], "Selected" ,
        col="red",  adj=c(0,0.5),cex=CEX)

#### pattern produced by proposed approach to (selective) SAMPLE (s) data
# try approach of Masters et al. model 3, ie fit additional modifier

d$s$year.born = (ds$YE - ds$AE ) - 1940
d$s$age.in.C = ds$AE - mean(entry.ages)
d$s$age.in.C.i.male = ds$age.in.C * ds$Male

head(ds)

proposed.s = glm(fewer.d ~ A + year.born + age.in.C + Male + A.M + age.in.C.i.male +
    offset(o), family=poisson, data=ds)

proposed.s  

# compare coefficient of A.M with its counterpart in unselected data

proposed.beta = proposed.s$coefficients[5:6]
HR.proposed.s = exp (proposed.beta[1] + proposed.beta[2]*(Ages-A.c) 
lines(Ages, HR.proposed.s , col="blue", lwd=1.5)
text(0.25+ Ages[length(Ages)], HR.proposed.s[length(Ages)] , "Proposed" ,
    adj=c(0,0.5),col="blue",cex=CEX)

# examine correlation of exposure*age and exposure*age.at.entry coefficients
# ie the 6,7 or 7,6 cell

round(cov2cor(summary(proposed.s)$cov.scaled),2)
# this would be more extreme if follow up were only to 2006

########## plot selectivity function

# segments(30,-0.6,100,-0.6, lwd=2, col="grey50")

Ae = c(50, 55, 60, 65, 70, 75)

COL="red"
y0 = -3.0; dy=1.75

for (ae in Ae ) {
    aa = ae + 0:(next.no.years-1)
    fx = fx.of.future.deaths.included(ae,1990,aa)
    points(aa,y0+fx*dy,cex=0.5,pch=19, col=COL)
    txt = toString(ae)
    if(ae == 50 | ae==75) txt = paste("Age at entry:",txt)
    #if(ae %in%
# c(50, 55, 60, 65, 70, 75) text(
# ae=0.2,y0+fx[1]*dy,txt,cex=CEX,adj=c(1.05,1),col=COL)
}

fx = c(0,1)
rect(40, y0, 85,y0+dy)

fx = seq(0.0, 1.0, 0.2)
for (f in fx){
  segments(xx, y0+f*dy, xx-0.5, y0+f*dy)
  text(xx-0.75, y0+f*dy, format(f,nsmall=1),
       adj=c(1,0.5), cex=CEX)
  segments(85, y0+f*dy, 85.5, y0+f*dy)
  text(85.75, y0+f*dy, format(1-f,nsmall=1),
       adj=c(0,0.5), cex=CEX)
}

text(xx-3.5, y0+0.5*dy,
     "Fraction of Deaths Included",
     adj=c(0.5,0), cex=CEX, srt=90)

text(88, y0+0.5*dy,
     "Fraction of Deaths Excluded",adj=c(0.5,1), cex=CEX, srt=90)

#COL="grey30"
text(32.75,y0+2.25,"B", adj=c(0,1),cex=CEX)
text(32.75,YLIM[2],"A", adj=c(0,0),cex=CEX)

for(xa in seq(40,85,5 )){
  segments(xa, y0, xa, y0-0.05)
  text(xa,y0-0.15,toString(xa),adj=c(0.5,1),cex=CEX)
}

text((40+85.5)/2,y0-0.55,"Attained Age, years",
     adj=c(0.5,1),cex=CEX)
# Web Figures 2 and 3, May 16 2016
# james.hanley@mcgill.ca

setwd("/Users/jameshanley/Dropbox/work/PAR-RT-TE/obesity/MortalityDatabase")

# Web Fig 2
# age-specific patterns in male:female HRs in entire populations

par( mfcol=c(2,3), mar = c(5,5,1.04,1.14))
country=c("Sweden", "Denmark", "Norway", "Switzerland", "France", "Finland")

for (c in 1:6) {
  fname = paste(country[c],"CohortDeathRates.txt",sep="")
  ds=read.table(fname,as.is=TRUE)
  ds=ds[, 1:4 ]
  head(ds)
  names(ds)=c("year.born","age","r.f","r.m")
  ds[,1]=as.numeric(ds[,1])
  ds[,2]=as.numeric(ds[,2])
  ds[,3]=as.numeric(ds[,3])
  ds[,4]=as.numeric(ds[,4])
  ds$year = ds$year.born + ds$age
  ds=ds[!is.na(ds$age) & !is.na(ds$r.m) ,]

  plot( c(40,100), c(0.75,4.5), ylab="Male:Female Hazard Ratio (HR)", xlab="Attained Age",col="white")

  if(c==1) text(40,4.5,paste("Web Fig. 2:",country[c]), adj=c(0,1),cex=1.1)
  if(c >1) text(40,4.5,paste("...",country[c]), adj=c(0,1),cex=1.1)
  text(95,4.5,"Birth Cohort", adj=c(1,1))
  abline(h=1:5)
  abline(h=1,lwd=3,col="red")
  COL=2
  for (yb in seq(1905,1945,10) ) {
    WHICH = ds$year.born == yb
    a = ds$age[ WHICH]
    HR = ds$r.m[ WHICH] / ds$r.f[WHICH ]
    print(cbind(a,HR))
    COL=COL+1
    lines(a,HR,col=COL)
    text(95, 4.35 -(yb-1900)/30 ,toString(yB),col=COL, adj=c(1,0.5))
  }
}

-----------------------------
# Web Fig 3 .. create 19 waves with healthy-male selectivity
# stay with ds for Finland from last loop (c=6)
# Yearly Births in Finland, to start each birth cohort

fname = paste(country[c],"Births.txt",sep="")
B=read.table(fname,as.is=TRUE)
B=B[1:29 ] ; B # Births , col 1 female, col 2 male [drop total]

# age 21 in 1986 , born 1965
# age 21 in 2004 , born 1983
# age 21 in 1986 , born 1902
# age 21 in 2004 , born 1920

par( mfcol=c(1,1),mar = c(5,5,0.5,0.5))

plot( c(40,100), c(0.75,4.5), ylab="Male:Female Hazard Ratio (HR)", xlab="Attained Age",col="white")

 text(65,4.5,paste("Web Fig. 3:",country[c]), adj=c(0,1),cex=1.1)
 text(95,4.5,"Birth Cohort", adj=c(1,1))
 abline(h=1:5)
 abline(h=1,lwd=3,col="grey60")
COL=2

29
for (yB in seq(1905,1945,10) ) {
  WHICH = (ds$year.born == yB)
  a = ds$age[ WHICH]
  HR = ds$r.m[ WHICH] / ds$r.f[WHICH ]
  #print(cbind(a,HR))
  COL=COL+1
  lines(a,HR,col=COL)
  text(95, 4.35 -(yB-1900)/30 ,toString(yB),col=COL, adj=c(1,0.5))
}
dta = NULL  # start the full pop. and sample datasets
sample=NULL;
TIMEUNIT = 5  # values of 100 or 1000 nearly eliminate selectivity
# use red curves to see how much attenuation results from lower values..
# 5 or 10 give more modest attenuation
fx=0.01  # intended 1% sampling fraction of each agegroup at each wave
# to make for a starker but simpler simulation ...
# 100% participation of women at all ages waves : fx unchanged
# progressive non-participation by older males and in later waves
# say by 0.5% per year of age from 30 onwards,
# & by 1% per calendar year from 1986 onwards
fx.30.1986 = fx ; slope.a = -0.01/2 ; slope.y = -0.01
fx.30.1986 = fx ; slope.a = -0.00001/2 ; slope.y = -0.00001  # almost none

fx.ageIn.yearIn.male = function(ageIn,yearIn) min(fx.30.1986, fx.30.1986 * exp( slope.a*(ageIn-30) + slope.y*(yearIn-1986) ) )
# so if 1/2 male 1/2 female, then averaged over both genders
# (1/2) of 100* + (1/2) of (100 to 100 - (85-30)*0.5% - (2004-1986)*1% = 54.5% )
# (1/2) of 100 + 1/2 of (ave of 100 and 54.5%) = 50 + 1/2 of 78 = 89%
# Turns out that a lower than target % of men participating will not by itself
# create bias. it is how much healthier they are than the non-participating males
# that matters, so I made the decreasing participation of men with inc. age 'almost none'
# BUT the healthy male participant phenomenon must increase with age at entry
# and here is where TIMEUNIT comes in..

healthy.male.mortality.ratio = function(ageIn) {
  fn = exp(- (1:5)/100) ^ (ageIn/TIMEUNIT) ;
  # Healthy male effects for 1st 5 yrs in survey
  # more pronounced at older ages
  return(fn[5:1])
}
lapply(40:84,healthy.male.mortality.ratio)  # have a look at what various TIMEUNIT values do
# can be drastic at older ages...
# 5 years as per Shaper article... but much more pronounced at older ages than they saw.
# now loop through each birth cohort ... y.b = year born
for (y.b in 1902:1982) {  # birth cohorts
  age = ds[ds$year.born==y.b,2]  # vector of future attained ages up to 2011,
  # where HMD death rates stop
  year = y.b + age  # their corresponding calendar years
  n = length(age)
  year.born = rep(y.b-1900,n)  # same birth.year for each of these, 1=1901
  i.male=rep(1,n) ;  # gender indicator for regression
  # male py's & deaths per year of attained age
  H.m = cumsum( ds[ds$year.born==y.b,4] )  # integral of hazard fn. males
  py = B[y.b-1900,2] * exp(-H.m) * 1  # survivors at each age x 1 y. = person years
d = c(B[y.b-1900,2],py[1:(n-1)]) - py[1:n]  # approx. deaths at each age  
# glm will warn multiple times about non-integer no.s of deaths but not relevant.  
# we are basically using EXPECTED numbers of deaths.... a non-issue.  

# Full Data -- concatenate onto previous cohorts.. just 5 variables per row  
# 1 row = 1 birth cohort 1 year of attained age, no. of py and deaths for that Lexis cell  

dta=rbind(dta, cbind(i.male, year.born, age,year,py,d) )  

# 19 waves of sample Data  
for( wave.year in 1986:2004) {  
  W = ( year >= wave.year )  # which years of followup  
  l.W = sum(1*W)  # how many years of followup  
  if (l.W > 0 ) {  
    year.in = rep(wave.year,l.W)  # repeat as many times as there are years of f-u.  
    age.in = wave.year - y.b  # age at entry  
    f.in = fx.ageIn.yearIn.male(age.in,year.in)  # fraction who enter  
    # f.in of males who participate is age.in and year.in dependent  
    # in these who do, a healthy.male effect for 1st 5 years,  
    # and critically this is age-at-entry dependent)  
    Healthy.Male.Ratio = healthy.male.mortality.ratio(age.in)  # **********  
    if(any(Healthy.Male.Ratio>1)) print(Healthy.Male.Ratio) # check  
    if(l.W <= 5) fraction.of.deaths = Healthy.Male.Ratio[1:l.W]  # **********  
    if(l.W > 5) fraction.of.deaths = c(Healthy.Male.Ratio,rep(1,l.W-5))  
  }  
  sample = rbind(sample,  
  cbind(year.in, i.male[W],year.born[W],age[W],year[W],  
  f.in * py[W],  
  f.in * d[W] * fraction.of.deaths ) )  
  # concatenate onto previous cohorts.. now 7 variables per row (year.in &  
  year.born)  
  # 1 row = 1 birth cohort 1 year of attained age, no. of py and deaths for Lexis cell  
}  

# females .. no selection other than 100*fx % of the living  
i.male=rep(0,n)  # gender indicator for regression  
H.f = cumsum( ds[ds$year.born==y.b,3] )  # as above, but with female death rates  
py = B[y.b-1900,1] * exp(-H.f)  # as above, but with female numbers  
d = c(B[y.b-1900,2],py[1:(n-1)]) - py[1:n]  # as above, but with female numbers  

dta=rbind(dta, cbind(i.male, year.born, age,year,py,d) )  # FULL POPULATION OF women  
for( wave.year in 1986:2004) {  
  W = ( year >= wave.year )  
  l.W = sum(1*l.W)  
  if (l.W > 0 ) {  
    year.in = rep(wave.year,l.W)  
    sample = rbind(sample,  
    cbind(year.in, i.male[W],year.born[W],age[W],year[W],  
    fx * py[W],  
    fx * d[W] ) )  
  }  
}  

AGELOWER = 40  # since m:f HR curve in POPULATION not linear all way down in ages,  
# restrict fit to linear part.. otherwise need quadratic terms in age for model 3  
# could set it to 40 but then need a quadratic model 3
# quadratic instead of linear will be fitted if it is < 60)

```r
# FULL Data
data.frame(dta) ; head(dta) ; summary(dta)
dta = dta[dta$year >= 1986 & dta$year <= 2006 & dta$age >= AGE.LOWER,]
head(dta);
sum(dta$d)
```

```r
FULL = aggregate(dta[,5:6],by=list(by=dta$i.male),sum)
FULL # curious to see how many deaths and PY
```

```r
# Sample Data
data.frame(sample) ; str(sample) ; head(sample) ; summary(sample)
names(sample)=c("year.in","i.male", "year.born","age","year","py","d")
sample = sample[sample$year >= 1986 & sample$year <= 2006 & sample$age >= AGE.LOWER,]
head(sample); str(sample);
sum(sample$d)
```

```r
SAMPLE = aggregate(sample[,6:7],by=list(by=sample$i.male),sum)
SAMPLE # way to see effect of amount of selectivity applied to males..
      # numbers of deaths in males should fall if use lower value of TIMEUNIT .
      # i.e more health-selectivity    ... a double check on simulated selection
      # graphs in black vs. red are another
```

```r
age.C=65 ; age.in.C = 40  # CENTERING
      # ideally, both should be weighed by where deaths are
      # and not where persons or person years are
      # but 50 taken close to in Masters et al.
      # or 60 if only fitting from 60 onwards
```

```r
if(AGE.LOWER >= 60) age.in.C = 60
if(AGE.LOWER < 60) age.in.C = 50
```

```r
# FULL
data.frame(dta)
dta$ageC = dta$age - age.C
data$ageC.i.male = dta$i.male * dta$ageC
```

```r
# SAMPLE
sample$ageC = sample$age - age.C
sample$ageC.i.male = sample$i.male * sample$ageC
```

```r
sample$age.in = sample$year.in - (1900+sample$year.born)
sample$age.in.C = sample$age.in - age.in.C
sample$age.in.C.i.male = sample$i.male * sample$age.in.C
```

```r
summary(dta)
sum(dta$d) ; sum(sample$d) # overall numbers of deaths
```

```r
##### pattern in FULL data ******************
# model 1 (use numbering close to that in Masters et al. )
fit = glm(d ~ ageC + year.born + i.male + offset(log(py)), family=poisson,data=dta)
summary(fit)
```

```r
# note 0.08 ( exp[0.08] = 1.08) for age, just as Gompertz found
# 'average' M:F HR of exp[.69] = approx 1.8
# allowing HR to vary with age (model 2, and lines up with Fig 1)
fit2 = glm(d ~ ageC + year.born + i.male + ageC.i.male + offset(log(py)),
family=poisson,data=dta)
summary(fit2)
```

```r
beta2 = fit2$coefficients[4:5]
```

```r
# M:F HR lower by 1-exp[-0.0238] = 2.5% per year of attained age
```
# POP plots show not quite linear, so add quadratic (q) term to bring closer to plot
# esp. if go down as far as age 40

data$ageC.sq = data$ageC^2
data$ageC.sq.i.male = data$i.male * data$ageC

fit2q = glm(d ~ ageC + ageC.sq + year.born + i.male + ageC.i.male + ageC.sq.i.male + offset(log(py)),
  family=poisson,data=data)

summary(fit2q)
beta2q = fit2q$coefficients[5:7]

AGES = AGE.LOWER:85 ; n.a=length(AGES) # ********** PLOT Fits
pattern2 = exp(beta2[1] + beta2[2] * ( AGES - age.C ) )
if(AGE.LOWER>= 60) lines(AGES,pattern2,lwd=1.5,col="red")

pattern2 = exp(beta2q[1] + beta2q[2] * (AGES-age.C) + beta2q[3]*(AGES-age.C)^2)
if(AGE.LOWER < 60) lines(AGES,pattern2q, lwd=1.5)

text(AGES[n.a],pattern2q[n.a],"Population",adj=c(-0.1,0.5))

######## pattern produced by proposed approach to (selective) SAMPLE (s) data

fits = glm(d ~ ageC + year.born + i.male + offset(log(py)), # s = sample
  family=poisson,data=sample)

summary(fits) # compare with FULL (neither model 1 is great)

summary(fit)

# model 2 .. ie effect modification by age

dd = aggregate(sample$d, by=list(a=sample$ageC,m=sample$i.male),sum)

yy = aggregate(sample$py, by=list(a=sample$ageC,m=sample$i.male),sum)

h = dd$x/yy$x

hr = h[66:130] / h[1:65]

aa = dd$a[1:65] + age.C

points(aa[1:46],hr[1:46], col="red",pch=19,cex=0.4)

fit2S = glm(d ~ ageC + year.born + i.male + ageC.i.male + offset(log(py)),
  family=poisson,data=sample)

summary(fit2S)
beta2S = fit2S$coefficients[4:5]

pattern2S = exp(beta2S[1] + beta2S[2] * ( AGES - age.C ) )
if(AGE.LOWER >= 60) lines(AGES,pattern2S,lwd=1.5,col="red")

# quadratic

sample$ageC.sq = sample$ageC^2

sample$ageC.sq.i.male = sample$i.male * sample$ageC

fit2Sq = glm(d ~ ageC + ageC.sq + year.born + i.male + ageC.i.male + ageC.sq.i.male + offset(log(py)),
  family=poisson,data=sample)

summary(fit2Sq)
beta2Sq = fit2Sq$coefficients[5:7]

pattern2Sq = exp(beta2Sq[1] + beta2Sq[2] * ( AGES - age.C ) + beta2Sq[3]*( AGES - age.C)^2)
# if(AGE.LOWER < 60) lines(AGES,pattern2Sq,lwd=1.5,col="red")

text(AGES[n.a],pattern2Sq[n.a],"Selective Sample,
  19 waves, 1996-2004",adj=c(1.25,1),col="red")

# if serious selectivity, should be lower than POP ones, and fall faster with age...

# try approach of Masters et al. model 3, ie fit additional modifier

fit3S = glm(d ~ ageC + year.born + age.in.C +
i.male + ageC.i.male +
age.in.C.i.male +
offset(log(py)), family=poisson, data=sample

summary(fit3S)
beta3S = fit3S$coefficients[5:6]
pattern3S = exp(beta3S[1] + beta3S[2] * ( AGES - age.C ) )
if(AGE.LOWER >= 60) lines(AGES,pattern3S,lwd=1.5,col="purple")
if(AGE.LOWER >= 60) text(AGES[n.a],min(pattern3S[n.a],43.5),"Approach of Masters et al. applied to Sample",adj=c(1,1),col="purple") # ********** PLOT Fits

# in case we want to fit down to lower ages, have a quadratic in age version of HRs
# and log-linear in age-at entry

fit3S.sq = glm(d ~ ageC + ageC.sq + year.born +
            i.male + ageC.i.male + ageC.sq.i.male +
age.in.C + age.in.C.i.male +
offset(log(py)), family=poisson, data=sample)

summary(fit3S.sq)
beta3S.sq = fit3S.sq$coefficients[5:7]
pattern3S.sq = exp(beta3S.sq[1]+beta3S.sq[2]*(AGES-age.C)+beta3S.sq[3]*(AGES - age.C)^2 )
if(AGE.LOWER < 60) lines(AGES,pattern3S.sq,lwd=1.5,col="purple")
if(AGE.LOWER < 60) text(AGES[n.a],min(pattern3S.n.a],4.25),"Approach of Masters et al. applied to Sample",adj=c(1,1),col="purple") # ********** PLOT Fits

text(40,4.35,paste("Models fitted to attained ages","toString(AGE.LOWER), "to 85"),
        font=3,cex=0.75, adj=c(0,1))

43.75