Estimation of Health Status Inequalities from Prevalence Data: A Risky Business

Alberto Palloni
and
Jason Thomas

Center for Demography and Ecology
Center for Demography and Health of Aging
University of Wisconsin-Madison

[0]Paper presented at the Population Association of America Meeting, Dallas, April 14-17, 2010.This study was supported by grants from the National Institute of Aging (R01 AG016209 and R37 AG025216) and Fogarty International Center (FIC) training program (5D43TW001586) to the Center for Demography and Ecology (CDE) and the Center for Demography of Health and Aging (CDHA), University of Wisconsin–Madison. CDE is funded by the NICHD Center Grant 5R24HD04783; CDHA is funded by the NIA Center Grant 5P30AG017266.

1 Introduction

Owing to the increased availability of cross sectional surveys current status data have gained popularity and are being increasingly used in a number of disciplines. This type of information and associated statistical tools are employed to draw inferences about the underlying incidence of a phenomenon and to identify the main determinants of its intensity and duration profile. Accurate inferences about incidence are generally made from data collection plans that follow observations over time thus permitting to identify more or less precisely the timing of occurrence of relevant events. However, implementation of longitudinal designs involves complex and expensive enterprises and are usually replaced by single wave cross-sectional surveys. Thus, the study of phenomena such as onset of illnesses, recovery from treatment, menopause, first intercourse, weaning, leaving home, first marriage and the like usually rely on information collected retrospectively in cross sectional surveys. Because retrospective recall of events and their timing is oftentimes inaccurate and unreliable, inferences about incidence stand on shaky ground and their worth as falsifying information is discounted. An alternative to retrospective recall of the timing of event is to use current status data, that is, information about the occurrence or non occurrence of a relevant event prior to a time marker such as a survey defined *ex-ante*. At an individual level the information is elicited from answers to questions that probe the experience of an event prior to the time marker. Such information is aggregated as prevalence data or the fraction of the population that experienced the event at the time of interview. Thus, for example, the proportion of mothers who at the time of the survey are still breastfeeding the most recently born child is used to make inferences about the timing of weaning (Grummer-Strawn, 1993). Similarly, the proportion of females at age x who are still single is used to identify characteristics of the timing of marriage (Hajnal, 1953). In the same vein, the occurrence or non occurrence of diabetes to an individual aged x at the time of the survey is material that, when properly handled, can yield useful insights about the proccesses that drive the incidence of diabetes.

However, inferences from current status data have an important drawback: they are not always robust to competing events that occur prior to the survey to some individuals who, as a consequence of them, are unable to provide information about their current status. For example, information about breastfeeding is not elicited from mothers whose most recently born child died before the survey. Individuals who migrate away from a household cannot provide information about their marital status. And members of a cohort who contract diabetes but die before the survey as a result of complicating factors, cannot provide information about their status. In such cases the validity of inferences about underlying incidence will depend strongly on whether or not the competing event occurs differentially among those who experience and those who do not experience the event of interest.

This paper discusses inconsistencies induced by the presence of competing events and proposes a simple way to reduce or eliminate them altogether. Al-

though these inconsistencies have been identified in the literature and are well-known among specialized researchers in the topic, they are conventionally dismissed as trivial or altogether ignored in empirical applications. What is new in this paper is the identification of conditions that lead to incosistency and the formalization of a simple correction procedure. The paper is organized as follows: Section 2 reviews the linkages between current status information and the underlying incidence function in a non formal, heuristic manner. In Section 3 we introduce population heterogeneity, indentify problems for inferences when there are non-ignorable competing events, and propose an adjustment procedure using maximum likelihood. Section 4 evaluates these procedures using Monte Carlo simulations. Section 5 summarizes results and concludes.

2 The algebra of current status

In this section we a provide informal description of current status data and associated statistics. It is not our intention to deliver a thorough review of these procedures. Kieding (1991) produced a very thorough survey of statistical perspectives undergirding current status methods. Diamond and MacDonald (1991) reviewed binary models for current status data, their relation to survival models and their demographic and epidemiological applications. Jewell and van der Laan (2002) offer an updated review of methods, extensions and applications. To fix ideas suppose we are interested in the occurrence of an event

E. Individuals in a population are characterized by a waiting (latent) time or duration d_i, defined as the elapsed time between the calendar time of onset of exposure t_{oi} and the calendar time of the occurrence of E, t_{ei}, as well as by a probability $P(E)$ that E will ever be experienced. We assume individuals are observed at an exact date, t_s, the date of a survey. The survey provides enough information to define an indicator variable $\xi_i = I_i(t_{oi} < t_{ei} < t_s) = 1$ if event takes place before survey and 0 otherwise. Ocassionally, but not always, the survey contain retrospective questions to elicit the timing of the event, t_{ei}. In general, however, this information is not collected or it is unrelaiable. If the event took place, that is, if $t_{ei} < t_s$ we have left censored data; if the event has not occurred, that is, if $t_{ei} > t_s$, we end up with right censored data. If the individual provides information on the date of the event, that is if t_{ei} is known (albeit with some error), we obtain partially left censored data.

As is frequently the case in demographic and epidemiological surveys, the age of individuals is the central measure of passage of time and generally we can translate the above time and duration indicators into an age metric. Let x_{toi} be the age of individual i at the time of onset of exposure,t_{oi} , x_{si} the age at the time of the survey, and x_{ei} the age at the time of event (if this took place) so that duration is $d_i = x_{si} - x_{ei}$. In those ocassions when d_i is known it

is subject to considerble noise and in what follows we will proceed as if it were unknown.

The foregoing information can be aggregated by age and by population subgroups. Assume that we have information on exact ages at the time of the survey, on the occurrence/non-occurrence of the event and on the presence/absence of a trait, Z, which the investigator believes exerts influence on the occurrence and timing of event E. The observable aggregated quantities will be denoted as $N(x, t_s, Z = 1)$ and $\overline{N}(x, t_s, Z = 1)$, the number of individuals with trait Z and aged x at the time of the survey who have and have not experienced E respectively. Analogous expressions apply for the number of individuals who do not possess the trait, namely, $N(x, t_s, Z = 0)$ and $\overline{N}(x, t_s, Z = 0)$. Whether or not we can observe all individuals who experienced E is wholly dependent on the process being studied. In most cases there is a set of competing censoring events $\{E_c, c = 1...k\}$ each characterized by a duration d_{cj} for individual j such that whenever $d_{cj} < (t_{sj} - t_{ei})$ we will have *no information whatsoever on these individuals*[1]. For example, in a cross-sectional survey of older people the researcher may have information on diabetes status for all pertinent ages x and no information at all on individuals who died whether or not they contracted diabetes before the survey. These processes are represented in Figure 1 as transitions between states, one characterized by the absence of event E another characterized by its presence, and a third representing events leading to unobserved individuald at time t_s. The notation in this figure makes explicit an important assumption we use throughout the paper, namely, that there are no cohort changes or, equivalently, that all processes leading to events of interest are stationary and do not depende on calendar time.

Let $\delta(y, Z = 1)$ be the instantaneous risk at age y of event E for individuals with trait Z, $\upsilon(y, Z = 1)$ the sum of instantaneous risks of competing censoring events among those with trait Z who experience E and, finally, $\mu(y, Z = 1)$ the instantaneous risk at age y of competing censoring events among those with trait Z who do not experience E. The investigator's interest is on the function $\delta(y)$ variously referred to as the "incidence" function, the "risk" function and the "hazard" function of event E.To simplify exposition we will assume that there is only one censoring competing event, mortality, and that the associated risk among those who experienced E is dependent on age but not on duration since the occurrence of E. All results we discuss in this paper differ only slightly if one assumes that $\upsilon(x, z)$ is also duration dependent.

The process represented in Figure 1 has been well-studied by Keiding (1991) and by authors interested in statistical inference from current status data (Diamond and McDonald, 1991; Keiding et al, 1995; Sun and Kalbfleish, 1993; Keiding et al., 1989; Kieding et al., 1996). But detailed attention to the problem generated by the existence of censoring competing risks has only recently been formally investigated (Jewel and Van der Laan, 2002). This paper rests on some of these new developments and proposes a tractable solution for empirical

[1] A competing censoring event Ec censors observations that may or may not have experienced event E in the sense that the calendar time of their occurrence to individual i, t_{ci}, is less than t_s.The individual is thus not observed at the time of the survey

estimation. In what follows we introduce the basic algebra of current status data and derive expressions in the case of homogeneous and non homogeneus risks. We first deal with the case when the population is assumed to be homogeneous and then when there is heterogenity with respect to a binary trait Z believed to have effects on the incidence of event E.

2.1 Case 1: Homogeneity of risks

Assume that there are no mortality differentials between those who experience E and those who do not, e.g $\upsilon(v) = \mu(v)$, that is, mortality risks for each subgroup are identical to some baseline mortality valid for the entire population. Assume also that onset of exposure is age 0 (birth). The probability of reaching age x at time t_s without experiencing event E is

$$\psi(x) = \exp(-\int_0^x (\mu(v) + \delta(v))dv) = \Phi(x) * \Lambda(x)$$

where $\Phi(x) = \exp(-\int_0^x \mu(v)dv)$ is the single decrement cumulated probability of surviving to age x in the absence of condition E and $\Lambda(x) = \exp(-\int_0^y \delta(v)\ dv)$ is the single decrement cumulated probability of avoiding event E in the absence of general mortality. If, as it happens with many conditions with adult onset, the process starts at arbitray age say $x_o > 0$, the above expressions hold with an origin shift.

Inferences about the process(es) leading to the occurrence of E focus on the density function $\delta(v)$ or any of the quantities defined by it, particulary the integrated hazard, namely, $(-\int_0^y \delta(v)\ dv)$. The observed data can be used to make inferences about $\delta(v)$ and the set of parameters on which it depends. Our purpose is to show that such inferences can only be made under quite restrictive assumption regarding competing censoring events.

If the number of entrances at origin $(t_s - x)$ years before is the stream $N(0,\ t_s - x)$, the expected number of individuals aged x who have not experienced E is given by

$$\overline{N}(x, t_s) = N(0, t_s - x)\ *\ \psi(x)$$

The probability of surviving to age x at time t for those who experience the E is

$$\Omega(x) = \int_0^x \exp(-\int_0^y (\mu(v) + \delta(v))dv) * \delta(y) * \exp(-\int_y^x \mu(v)dv)dy$$

or

$$\Omega(x) = \exp(-\int_0^x \mu(v)dv) * \int_0^x \delta(y) * (\exp - \int_0^y \delta(v)dv)dy = \Phi(x) * (1 - \Lambda(x))$$

The expected number of surviving individuals aged x and who experience E before t_s, $(t_{ei} < t_{si})$, is

$$N(x, t_s) = N(0, t_s - x) \ * \ \Phi(x) * (1 - \Lambda(x))$$

the neat factorization being possible only because $\mu(x) = \upsilon(x)$.In this case the observed proportion of individuals who experienced E by age x,e.g. the prevalence of E at age x, is an empirical estimate of the probability of experiencing event E before age x (see also Kieding, 1991):

$$p(x,t) = \frac{N(x,t)}{N(x,t)+\overline{N}(x,t)} = \frac{\Phi(x)*(1-\Lambda(x))}{\Phi(x)*(1-\Lambda(x))+\Phi(x)\Lambda(x)} = (1 - \Lambda(x))$$

and, conversely, the observed proportion who have not experienced E, $\overline{p}(x,t)$, is an estimate of $\Lambda(x)$, the single decrement probability of not experiencing E. Thus, under risk homogeneity, the observed prevalence rates at ages x (current status observable) provide sufficient information to generate Nelson-Aalen type estimates of the integrated hazard of event E and, under minimal regularity conditions, estimates of the risk or intensity of event E, the target quantity (Kieding, 1991).

2.2 Case II: Heterogenity of risks

Suppose that $\mu(x) \neq \upsilon(x)$. $\mu(x)$ is now a baseline mortality risk for the general population. The expression for the function $\Omega(x,t)$ becomes

$$\Omega(x) = \int_0^x \exp(-\int_0^y (\mu(v) + \delta(v))dv) * \delta(y) * \exp(-\int_y^x \upsilon(v)dv)\, dy$$

Further simplification can be achieved if, without loss of generality, we assume that $\mu(x)$ and $\upsilon(x)$ are linked through a function $g(\theta)$, e.g., $\upsilon(x) = \mu(x)g(\theta)$ where θ is an arbitrary parameter determining the difference in mortality risks. The expression for $\Omega(x)$ becomes

$$\Omega(x) = \exp\,(-\int_0^x \mu(v)dv) * \{\int_0^x \delta(y) * \exp\,(-\int_0^y \delta(v)dv) \ * \\ * \exp(-\int_y^x (\mu(v)(g(\theta) - 1))dv\ dy\}$$

or

$$\Omega(x) = \Phi(x) * \{\int_0^x \varphi(y) * \exp\,(-g(\theta)\int_y^x \mu(v)dv)dy\}$$

where $\varphi(y)$ is the density of the waiting times to experience event E at age y. To gain more transparency we use the mean value theorem and re-express $\Omega(x)$ as :

$$\Omega(x) = \Phi(x) * (1 - \Lambda(x)) * \alpha(x, \widetilde{y}_x)$$

where $\alpha(x, \widetilde{y}_x) = \exp(-g(\theta)\int_{\widetilde{y}_x}^x \mu(v)dv)$ and $0 < \widetilde{y}_x < x$, $\widetilde{y}_x$ is an implicit function of $\varphi(y)$, θ, and $\mu(v)$. Thus the observed proportion who experience event E by age x is

$$p(x,t_s) = \frac{N(x,t)}{N(x,t)+\overline{N}(x,t)} = \frac{(1-\Lambda(x))\alpha(x,\widetilde{y}_x)}{(1-\Lambda(x))*\alpha(x,\widetilde{y}_x)+\Lambda(x)} \leq (1-\Lambda(x))$$

whereas the observed proportion who survived without experiencing E is

$$\overline{p}(x,t_s) = \frac{\overline{N}(x,t)}{N(x,t)+\overline{N}(x,t)} = \frac{\Lambda(x)}{(1-\Lambda(x))*\alpha(x,\widetilde{y}_x)+\Lambda(x)} \geq \Lambda(x)$$

When $g(\theta) = 0$ and $\alpha(x,\widetilde{y}_x) = 1$ we are back to a situation of risk homogeneity. When $g(\theta) \neq 0$, that is, when individuals who experience E are exposed to different mortality risks than the general population, the value of $\alpha(x,\widetilde{y}_x)$ can be smaller or larger than 1 and, as a consequence, the observed proportions who experience event E do not provide enough information to retrieve estimates of the hazard or integrated hazard associated with the event. If E is a disease, such as diabetes, it is likely that $g(\theta) > 1$ and observed prevalence will underestimate the quantity of interest, $(1-\Lambda(x))$. Under other circumstances, $g(\theta) < 1$ and the observed prevalence will overestimate the probability of experiencig the event before age x. For large samples and when $|\alpha(x,\widetilde{y}_x)|$ is close to 1 the bias in $\overline{p}(x,t_s)$ is approximately equal to $(1-\alpha(x,\widetilde{y}_x))(\Lambda(x)*(1-\Lambda(x)))$.This expression attains a maximum at age x_{max} when $\Lambda(x_{\max}) = .5$. Since $\alpha(x,\widetilde{y}_x)$ decrease with age, the error in the observed prevalence rates will increase at least up to x_{max}. If $\Lambda(\infty) < .50$ the observer will be fooled into believing that those who are *older are less likely to experience diabetes than those who are younger*. Take as illustration the case of diabetes: mortality among diabetics is likely to be higher than among those without it, e.g., $g(\theta) > 1$. Figure 2 displays the magnitude of relative bias in four different scenarios resembling what one would get in such cases[2]. The curves in Figure 2 represent the proportionate errors in the estimate of $\Lambda(x)$. Since the errors increase with age, the observed age pattern of prevalences contains a downward bias that worsens with age and the observer will interpret the figures as suggesting that the incidence of the phenomena has been less intense for the older cohorts.

3 Population heterogeneity

Our next step is to relax the assumption of population homogeneity, assume that we observe individuals with and without trait Z and that we wish to make

[2] The values of $\widetilde{y}$ were set to be 45, 50, 55 at ages 60,65 and 70 and 60 at ages above 70. Two sets of values of Λ_1 and Λ_2 were used. They correspond to the conditional survival curves from age 55 onward (in intervals of 5 years) in the Coale-Demeny system of life tables (Model West, females) with life expectancies equal to 78 and 80 respectively. Λ_1 atains a maximum value of .581 ate age 100 and Λ_2 attains a maximum of .756. Thus the incidence regime is more punitive in the first set of values, Λ_1.Finally, we used two alternative values $\alpha, \alpha_1$and α_2, calculated from the same life tables with $g(\theta) = 1.50$ (mortality differential of 50 percent). The four curves in Figure 2 display the relative errors in the estimate of Λ_2 and Λ_2 and correspond to the combinations of Λ_1 with α_1 (relerror1) and α_2 (relerror2), and Λ_2 with α_1 (relerror3)and α_2 (relerror4) respectively.

inferences about the effect of Z on the incidence of event E. We use current status information and proceed to compare the prevalence of the condition at various ages in subgroups with and without Z. We know that in the case of homogeneity of risks the observed prevalences in each subgroup is sufficient to obtain unbiased estimators of the true single decrement survival probabilities of experiencing E . It therefore must be the case that a contrast of prevalence rates across subgroups yields an unbiased estimator of the effect of trait Z on the risk of of contracting the condition. This will not be the case under a regime of risk heterogeneity in one or both subgroups.

3.1 Estimates of effects of covariates: informal approach and approximations

Suppose the effects of trait Z on the risk of experiencing event E can be represented by a proportional hazard model, $\delta(x, Z = 1) = \exp(\lambda) * \delta(x, Z = 0)$. In this case the values of the single decrement probabilities of not experiencing E in the two subgroup should be related as

$$\Lambda(x, Z = 1) = \Lambda(x, Z = 0)^{\exp(\lambda)}$$

so that the log-log transforms of the single decrement probabilities of not experiencing E are related linearly to each other with an offset equal to $\exp(\lambda)$.If the assumption of risk homogeneity is accurate an estimate of $\exp(\lambda)$ can be retrieved from prevalence data. In fact the ratio of prevalence rates in the two subgroups at age x is given by:

$$O^T(x) = \overline{p}(x, Z = 1)/\overline{p}(x, Z = 0) = \Lambda(x, Z = 1)/\Lambda(x, Z = 0) =$$
$$= \Lambda(x, Z = 0)^{\exp(\lambda)-1}$$

whereas under risk heterogenity the ratios of observed proportions are more complex functions of $\exp(\lambda)$ and of the quantities $\alpha(x, Z = 1)$ and $\alpha(x, Z = 0)$:

$$O^o(x) = \overline{p}(x, Z = 1)/\overline{p}(x, Z = 0) = \Lambda(x, Z = 0)^{\exp(\lambda)-1} *$$
$$* \frac{\Lambda(x,Z=0)+(1-\Lambda(x,Z=0)*\alpha(x,Z=0)}{\Lambda(x,Z=1)+(1-\Lambda(x,Z=1)*\alpha(x,Z=1)}$$

The exact magnitude of the bias in a hazard model-based estimate depends on the relative magnitudes of $\alpha's$ and $\Lambda's$.To provide an idea of the size of the error Figure 3 displays age-specific estimates of λ (from observed prevalence by age) and for four combinations of $\alpha(x, Z = 1)$, $\Lambda(x, Z = 1)$ and a fixed value

of $\lambda = .5$. Note that the estimates can contain substantial (negative) errors so much so that in some cases is even improperly signed[3].

In summary, inferences about the underlying incidence of an event and of the size of effects will contain sizeable biases when there is population heterogeneity, differential attrition due to competing events among those who experience and do not experience E and, finally, differences in the attrition processes across subpopulations. The approximations illustrated in Figures 2 and 3 suggest that the use of prevalence data while igoring heterogeneity of competing risks leads to understimation of cumulated incidence. In general the bias will get worse as duration from the time of initiation of the event increases. A possible illustration of this bias occurs when examining prevalence rates of of diabetes by age: in many cases the rates tend to bend downwards with age, as Figure 2 suggest they would. This does not necessarily mean that incidence among older people is lower than among the younger ones. Similarly, data of prevalence of diabetes by levels of education shows a notorious regularity: in many cases the curve for the least educated converges and sometimes crosses over the curve representing the prevalence rates of those with higher levels of education. As shown by Figure 3 this could simply be a result of heterogeneity of mortality between diabetics and non diabetics.

3.2 A likelihood approach to adjustments

The likelihood of a sample of current status observations when there is no mortality differentials among those who do and do not experience the event is

$$\mathcal{L} = \prod_{i=1}^{N} \{ \exp(-I^m(x_i)(1-\exp(-I^E(x_i)))\}^{Y_i} \{ \exp(-(I^m(x_i) + I^E(x_i)))\}^{(1-Y_i)} =$$
$$\prod_{i=1}^{N} \exp(-I^m(x_i)) * \{(1-\exp(-I^E(x_i)))\}^{Y_i} \{ \exp(-(I^E(x_i)))\}^{(1-Y_i)}$$

where $I^m(x_i)$ and $I^E(x_i))$ are, respectively, the integrated hazards from 0 to x_i for mortality and event E and Y_i =1 if an individual i observed at time t_s experiences the event and 0 if the individuals is observed but did not experience the event. Since $\mathcal{L}$ is an unconditional likelihood we should standardize by the probability of surviving up to t_s and upon doing this the term $\exp(-I^m(x_i))$ drops out of the expression. Again the neat factorization of the likelihood is possible because the baseline mortality function carries no information about the incidence of E or baseline mortality is ignorable with respect to E [4].

[3] The values of Λ and α and the four combinations created with them are the same as for Figure 2 (see previous footnote)

[4] The likelihood advocated by Diamond and MacDonald is one where mortality (or any other type of competing attrition) is treated as ignorable

When there is differential mortality between those who experience and those who do not experience E the likelihood is given by:

$$\mathcal{L} = \prod_{i=1}^{N} \exp\left(-I^m(x_i)\{\exp\left(-I^E(x_i)\right)\}^{(1-Y_i)}\right.$$
$$\{(1-\exp\left(-I^E(x_i)\right))\}^{Y_i}\{\int_0^{x_i} f^e_{x_i}(y)\exp\left(-g(\theta)I^m(y,x_i)\right)dy\}^{Y_i}$$

where $f^e(y)$ is the conditional density of $(d_i|d_i < x_i)$ defined by
$(\delta(y)dy/\int_o^{x_i}\delta(y)dy)$ and $I^m(y,x_i)$ is the integrated hazard from age y to x_i. Assuming ignorability of the baseline mortality risks, the likelihood can be written as

$$\mathcal{L} \propto \prod_{i=1}^{N}\{\exp\left(-I^E(x_i)\right)\}^{(1-Y_i)}\{(1-\exp\left(-I^E(x_i)\right))\}^{Y_i}$$
$$\{\int_0^{x_i} f^e_{x_i}(y)\exp\left(-g(\theta)I^m(y,x_i)\right)dy\}^{Y_i}$$

The inner integral in the likelihood is the (conditional) expected value of the function $\exp(-g(\theta)I^m(y,x_i))$ which, using the delta method, can be approximated as:

$$\int_0^{x_i} f^e_{x_i}(y)\exp(-g(\theta)I^m(y,x_i))dy \cong exp(-Ex(g(\theta)I^m(y,x_i))) \cong$$

$$\exp(-g(\theta)I^m(Ex(y,x_i))) = \exp((-g(\theta)I^m((\widetilde{y}_{x_i},x_i))) = \varphi(\theta;\widetilde{y}_{x_i})$$

where Ex stands for expectancy and all expectations are with respect to the density $f^e_{x_i}(y)$ and the quantity $\widetilde{y}_{x_i}$ is the mean age at which Ex is experienced conditional on experiencing it before age x_i.The likelihood is now

$$\mathcal{L}\propto \prod_{i=1}^{N}\{\exp(-I^E(x_i))\}^{(1-Y_i)}\{(1-\exp(-I^E(x_i)))\}^{Y_i}\{\varphi(\theta;\widetilde{y}_{x_i})\}^{Y_i} =$$
$$\prod_{i=1}^{N}\{(1-p_{x_i})\}^{(1-Y_i)}\{p_{x_i}\}^{Y_i}\{\varphi(\theta;\widetilde{y}_{x_i})\}^{Y_i}$$

In conventional current status models the researcher can specify the nature of $\delta(y)$ (and therefore of p_{x_i}),make it dependent on a vector of parameters γ and then use standard likelihood procedures to obtain estimates of γ. However, in the presence of heterogeneous risks the strategy is no longer feasible. A further simplification is possible if we condition on survival to t_s and work on the conditional likelihood

$$\mathcal{L} \propto \prod_{i=1}^{N}\{(1-p_{x_i})/((1-p_{x_i})+p_{x_i}\varphi(\theta;\widetilde{y}_{x_i}))\}^{(1-Y_i)}$$

$$\{p_{x_i}\varphi(\theta;\widetilde{y}_{x_i})/((1-p_{x_i})+p_{x_i}\varphi(\theta;\widetilde{y}_{x_i}))\}^{Y_i}$$

This expression can be generalized to the case of heterogeneous populations: there will be covariates contained in a vector Z with effects on one or more of the parameters γ determining the incidence of the event. A transparent way of making $\mathcal{L}$ tractable is to assume that the log of the waiting times are logistic so that the odds can be expressed as an exponential of a linear combination of parameters

$$p_{x_i}/(1-p_{x_i}) = \exp(\beta Z)$$

where Z is a vector of covariates and β a vector of effects. Replacing this in the conditional likelihood we get after simplification

$$\mathcal{L} \propto \prod_{i=1}^{N}\{1/(1+\exp(\beta Z_i+\varphi(\theta;\widetilde{y}_{x_i})))\}^{(1-Y_i)}$$
$$\{\exp(\beta Z_i+\varphi(\theta;\widetilde{y}_{x_i})/(1+\exp(\beta Z_i+\varphi(\theta;\widetilde{y}_{x_i})))\}^{Y_i}$$

that is, the likelihood of a conventional logistic model with the set of covariates expanded to include $\varphi(\theta;\widetilde{y}_{x_i},x_i)$.

3.3 What should $\varphi(\theta;\widetilde{y}_{x_i})$ *be*?

Leaving aside for the moment possible departures from the assumption of log logistic waiting times, it is of relevance to examine the nature of the quantity $\varphi(\theta;\widetilde{y}_{x_i})$. It stands for the expectd value of the probability of surviving from the age at which E occurs up to age x_i and we are replacing it for the probability of surviving from the expected age at which E occurs to age x_i. It is the predicted probability of being 'eligible' for the survey after event E occurs and is explicitly defined for those who experience the event and implicitly for those who did not. The value of $\widetilde{y}_{x_i}$ for individual i can be calculated exactly only if we know what we are trying to estimate, namely, the incidence curve and its individuals determinants.However, it can be approximated from retrospective (possibly erroneous) information about the occurrence of E or from known incidence curves of E in the population under study or similar populations. In all these cases we can use a known mortality function for the total population as an approximation for $\mu(y)$.

Assume that $\varphi(\theta;\widetilde{y}_{x_i}) = \exp(-\theta I^m(\widetilde{y}_{x_i},x_i)), e.g$ the mortality risk of those who experience E is $(1+\theta)$ times as high as the mortality of those who do not. In this case introducing $\ln(\varphi(\theta;\widetilde{y}_{x_i})) = -\theta I^m(\widetilde{y}_{x_i},x_i)$ as a regressor in the logistic model leads to an estimate of θ. If the differentials in mortality apply to both subgroups created by Z, we can introduce two terms of the form $ln(\varphi(\theta;\widetilde{y}_{x_i}))$, one for each subgroup and associated with the scaling factors θ_1 and θ_2.Separate identification of these scaling factors may not always be possible and will depend on the precision with which the ages $\widetilde{y}_{x_i}$ are estimated and the degree to which

they differ in the two groups for a given age x. Even if the waiting times for E are drawn from identical distributions, the two subgroups will differ on the probabilities of experiencing E over a lifetime and this difference is enough to generate discrepancies in the values of $\widetilde{y}_{x_i}$ in each of the groups. In teh worst case scenario one can always estimate the difference $\theta_1 - \theta_2$.The most important point is that the term $\varphi(\theta; \widetilde{y}_{x_i})$ is an effective control to get unbiased estimates of the effects of Z and the parameter associated with it is ancillary and not of central interest.

When estimated as suggested above, the function $\varphi(\theta; \widetilde{y}_{x_i})$ will be invariant by age, that is, different individuals with the same age will be assigned the same value of the function, independently of covariates or of their status with respect to E. If age were to be measured exactly this overlap is less serious but in every-day cases we conventionally use months or years thus lumping individuals in coarse age groups.In these cases there will be a set of individuals indistinguishable with respect to age and therefore with the same value of the function. As a result of this inaccuracy the variability necessry for this function to operate as an adjustment factor will be obtained only as a by-product of its variability across the set of ages spanned by the sample.

If one thinks of surviving to the time of the survey as a selection process that enables the investigators to examine an outome, e.g. the occurrence/non occurrence of E in the past, then $\varphi(\theta; \widetilde{y}_{x_i})$ can be seen as performing the same role as does the adjustment for selection suggested by Heckman (1979). In clinical trials, where individuals are exposed to a treatment to assess the benefits of a therapy, only a subset of the baseline sample is offered the tratment as some individuals attrite before receiving it. In these cases one could use the same adjustment factor to correct the estimated effect of the treatment on the patients who are exposed to it.

4 Monte Carlo simulation

To assess the magnitude of the biases associated with current status estimators under conditions of population and risk heterogeneity and to evaluate the performance of the proposed adjustment, we simulate a population with two subgroups, high and low education, where the event E is diabetes, there is differential mortality between diabetic and non diabetic, and the lifetime risk of diabetes differs between the subgroups though the waiting times for diabetes are from a log normal (mean = 25 years, sd = 10 years) for both subgroups. Figure 4 shows the prevalence rates by age that would be observed in the absence of mortality. These are the prevalence rates that lead to unbiased estimates of the risks of E. We introduce general mortality, $\mu(x)$, as a Gompertz with a level parameter (exp(-8) = 0.0003) and a shape parameter 0.10. Mortality differentials between diabetics and nondiabetics are parameterized by changing the level parameters.

At the outset (age 30) there is an equal number of individuals in the two subgroups, but as age increases the proportion in the low education subgroup is higher reflecting differential growth rates by education. In particular, we assume that the rates of increase are constant and equal to .005 for high education and .001 for low education. In each subgroup we simulate 51 cohorts to represent populations at ages 50, 51,...,100 in a cross section. We assume that the minimum age at onset of diabetes is 30. Individuals in each cohort are assigned a waiting time to death, W1, between 0 and 70 according to the chosen Gompertz baseline mortality rate (see above). We then divide each cohort into two sets, one consisting of individuals who contract diabetes over their lifetime (between ages 30 and 100) and the other consisting of individuals who remain diabetes-free. These fractions are initially chosen so that the lifetime probability of developing diabetes is .20 in the high education group and .40 in the low education group. Individuals in the diabetic pool are randomly allocated a waiting time for the onset of diabetes, W2, using a draw from a log normal distribution with a mean (sd) of 25 (10) years. If W2 < W1 the individual is tagged as diabetic from age (30 + W2) onward. If W1 < W2, the individual is tagged as alive with no diabetes for all ages between 30 and (30 + W1) and dead for older ages. For those in the pool of diabetics we draw a waiting time to death starting from the age of onset of diabetes (30 + W2). This draw, W3, is from the Gompertz curve chosen to represent mortality among diabetics. This individual will be tagged as alive with no diabetes for all ages from 30 up (30 + W2), as alive and diabetic for all ages between (30 + W2) and (30 + W2 + W3), and as dead at higher ages. Finally, cohorts are joined so that one of them represents the population aged 50, another the population aged 51 and so on. Measures of age-specific prevalence are calculated. If one simulates the same cohorts but removing mortality risks, we obtain the underlying age-specific prevalences, that is, those corresponding to the underlying incidence function. To assess the effects of education we estimate logistic regressions on the age-specific prevalence rates using education as a dummy variable. The coefficient of education is a measure of the influence of education on the underlying incidence function and, in turn, corresponds to the contrast between the fractions that ultimately develop diabetes in both subgroups. We compare the estimates of the education coefficient across simulations with and without the risk of mortality to evaluate the magnitude of the biases and the performance of the adjustment procedure. The estimates will vary slightly from one simulation to the next since we are sampling from distributions of waiting times and the presence of diabetes, so we present results for 25 runs of each simulation to show the stochastic variability[5].

We begin by demonstrating the conditions under which differentials in mortality associated with diabetes result in biased estimates of the effect of education, as measured by the coefficient for the education (dummy) variable in what will be referred to as the naive model. Figure 5 shows the coefficient estimates, across 25 runs of each simulation, with each panel corresponding to different

[5] In all cases the logistic regressions include a control for age.

sizes of the mortality differentials associated with diabetes, and in each case baseline mortality is the same for the low and high education groups. In the scenario shown in panel (a) the mortality differential for the low education group (bottom axis) is equal to the mortality differential in the high education group (top axis). One can show algebraically that although this is not a situation of risk homogeneity, the estimate of effects will be very close to the true one, the difference between them being in most cases trivial. Bu empirical situations where risk heterogeneity is identical in both subgroups may be rare. If there is no mortality differential in the high education group, then there is a downward bias in the coefficient that increases in magnitude with the size of the mortality differential among the low education group, as shown in panel (b). Similarly, panel (c) shows the bias corresponding to the situation where there is a mortality differential among the high education group, but a larger differential among the low education group. In the final scenario shown in panel (d) the mortality differential in the high education group is proportional to the mortality differential in the low education group. Under these conditions the magnitude of the downward bias does not depend on the size of the mortality differential in the low education group. Similar results to those shown in Figure 5 are obtained when the populations are simulated with a higher level of baseline mortality for the low education group relative to the high education group.

We now evaluate the performance of the proposed adjustment procedure by examining the bias (i.e. coefficient estimated without mortality - coefficient estimated with mortality) of the naive and adjusted estimates across a combination of mortality differentials for the low and high education groups. The results, shown in Figure 6 , indicate a substantial reduction in the bias when the adjustment procedure is used, a finding which is robust to the sizes of the differentials among each education group. In the simulations upon which these results are based, the baseline mortality is the same for the low and high education group, but the findings are similar when we simulate populations with a higher level of baseline mortality for the low education group. We also find that the adjustment procedure performs fairly well under conditions where the mean waiting times to diabetes is different for the low and high education groups. The results up to this point have been obtained from simulations with a mean waiting time of 25 years and a s.d. of 10 years. Figure 7 shows the bias of the naive and adjusted estimates when the mean waiting time for the low education group is 5 years earlier than that of the high education group (20 versus 25 years), there is only a mortality differential (associated with diabetes) among the low education group, and baseline mortality is the same for everyone in the population. As the bias of the naive estimate increases with the mortality differential in the low education group, the adjusted estimate has very little bias even at excessively high levels. Although this robustness check is reassuring, more sensitivity analysis are warranted with respect to distributional assumptions and larger differences in the distribution of waiting times across groups.

In some situations information may also be available on the time at which the event of interest occurred, such as a self-reported age at which the individual was diagnosed with diabetes. This information can be used to calculate

the (conditional) mean age at which the event occurred (given that the event has been observed), $\tilde{y}_{x_i}$, and used to evaluate adjustment factor $\varphi(\theta, \tilde{y}_{x_i})$ for the logistic regression. We explore the performance of the adjustment procedure under these and other circumstances concerning the quality of the data. Figure 8 shows the bias in the naive and adjusted estimates of the coefficient for education in the logistic regression model for various scenarios.[6] The scenario shown in panel (a) is the same as those seen earlier which do not use any information on the time at which the individual was diagnosed with diabetes, and is included as a reference point. Panel (b) contains the results based on calculating the adjustment factor using the mean age at which individuals developed diabetes, a situation where the self-reported information is exact. In reality there is likely to be some error in the self-reported information, which we mimic in the third scenario, shown in panel (c), by adding a random draw from a normal distribution with mean of 0 and standard deviation of 2 years to the age when the individual developed diabetes (among those with diabetes).[7] Finally, we expanded on the previous scenario by adding a **systematic bias** that is positively associated with age so that older individuals are more likely to report an age at diabetes closer to their age at observation, the results for which are shown in panel (d). In each case where the self-reported information is used (i.e. panels (b), (c), and (d)), the adjusted estimates result in a negative bias (i.e. estimated coefficients that are greater than the actual value). There is virtually no impact when the self-reports contain random noise, but a systematic bias that is positively associated with age does appear to *reduce* the bias, which suggests that the (conditional) means from the self-reports are lower than those obtained from using the (log normal) density function. Thus, it appears that the performance of the proposed adjustment is sensitive to the value of the conditional mean age when the event occurred, $\tilde{y}_{x_i}$, and that differences between the assumed distribution of waiting times and the distribution that is generating the data may also hamper the performance of the adjustment procedure. In each case, however, the adjusted estimates have equal or less bias than the estimates from the naive model.

[6] In these simulations, there is only a mortality differential among the low education group, the baseline mortality is the same for everyone, and the distribution of waiting times to diabetes is also the same across education groups.

[7] If adding this random error resulted in an age at diabetes that was greater than the age at observation, then the age at observation was used as the age at diabetes.

5 Summary and conclusion

By and large, conventional current status analysis gives short shrift to potential errors that arise when risk of attrition prior to the time marker at which individuals' status is assessed, t_s, may vary as a function of the occurrence/non-occurrence of the event of interest. Through suitable approximations and simulations we show that even under mild conditions governing risk heterogeneity the biases can be substantial and could lead to misleading inferences about the time profile of the underlying risks and/or about the effects of covariates. The adjustment procedure we propose is simple, can be deployed with little effort and with minimal knowledge about the age pattern of the baseline risk of the competing event attrition. Simulations show that the adjustment performs (a) much better than the naive estimate, (b) is robust to the precise function governing the incidence of the event of interest and (c) is quite insensitve to measurement errors that may surface if the researcher uses self-reported information on time of occurrence for the calculation of the adjustment factor.

Future research should proceed along three different routes. The first is to investigate the asymptotic properties of the estimator suggested here. While in the case of a logistc funtions these are well understood, it is not so for other equally plausible functional forms. The second is to assess the robustness of the adjustment to an inaccurate rendition of the baseline for the competing risk. The integrated hazard on which the adjustment factor rests cannot be calculated without knoweledge of this baseline hazard. In cases when the competing risk is adult mortality, this may not be so difficult as what matters for the adjustment in these cases is to identify correctly the curvature of the hazard over the span of ages of interest, not its level. But in other applaications it may not be so clear what should the baseline hazard look like, let alone what its approximate curvature may be within a particular range of durations. The third route of research is to assess the performance of the proposed adjustment in a broad array of empirical cases and determine the extent to which they lead to different inferences.

REFERENCES

1. Diamond, I.D. and J.McDonald, 1991. "The analysis of current status data" In T.J. Trussell, R.Hankinson and J.Tilton (eds). *Demographic Applications of Event History Analysis*, Oxford: Oxford University Press, pp. 231-252

2. Grummer-Strawn, L.M. 1993. "Regression analysis of current status data: An application to beastfeeding" *Journal of the American Statistical Association* 881:758-765

3. Hajnal, J. 1953. "Age at marriage and proportions marrying" *Population Studies* 72(2):111-136

4. Heckman, J. 1979. "Sample selection bias as a aspecification error" *Econometrica* 47: 153-161

5. Jewel, N.P. and van der Laan, M.J., 2002. " Current status data: review, recent developments and open problems". *University of California-Berkeley Division of Biostatistics Working Paper Series*, paper No113

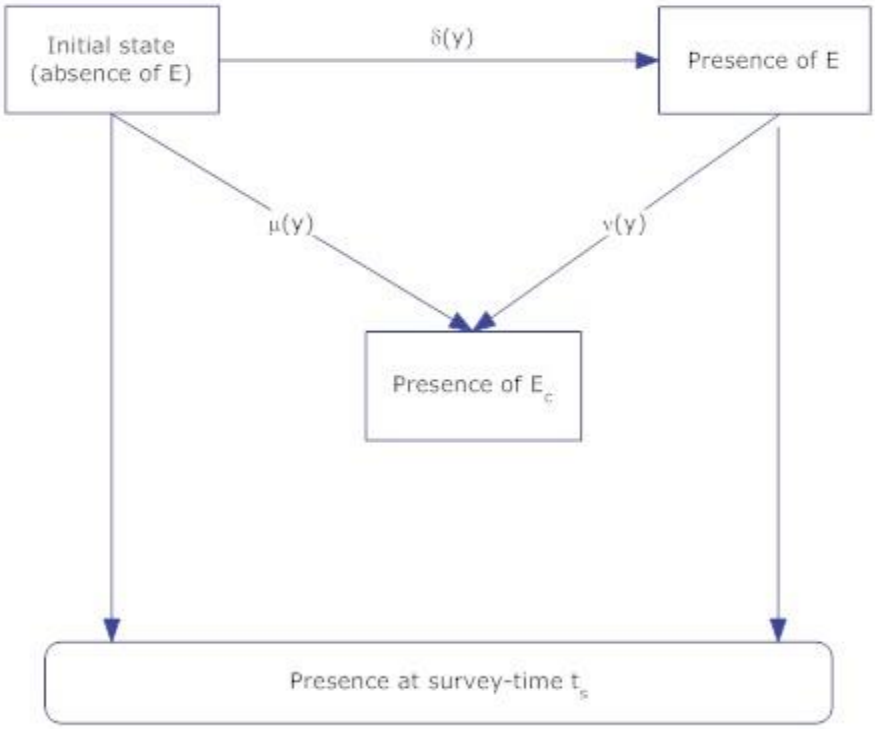
6. Kieding, N., Holst, C and Ansers, G. 1989. "Retrospective estimation of diabetes incidecne from information in a prevalent population and historical mortality" *American Journal of Epidemiology* Vol 130(3): 588-600

7. Keiding, N. 1991. "Age-specific Incidence and prevalence: a statistical perspective" *Journal of the Royal Statistical Society.* Series A, Vol 154(3): 371-342

8. Kieding, N., Begtrup, K., Scheike, T.H., Hasibeder, G. 1996. "Estimation from current-status data in continuous time" *Lifetime Data Analysis* 2:119-129

9. Sun, J. and J.D. Kalbfleish, 1993. "The analysis of current status data on point processes" *Journal of the American Statistical Association*, Vol 88(424): 1449-1454

Figure 1: Latent transitions between states in current status information with competing censoring events



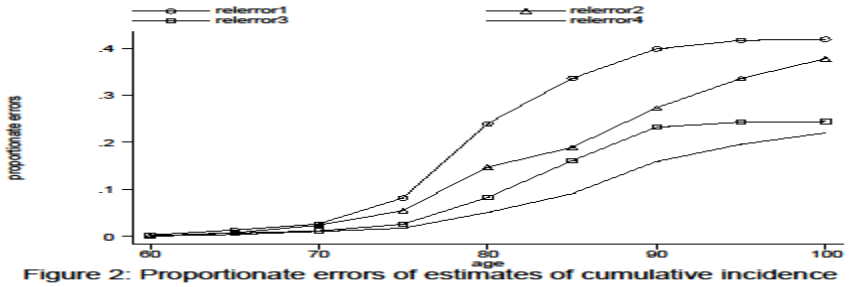


Figure 2: Proportionate errors of estimates of cumulative incidence

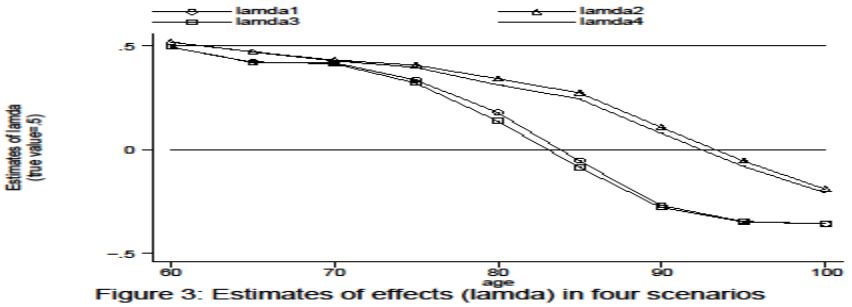


Figure 3: Estimates of effects (lamda) in four scenarios

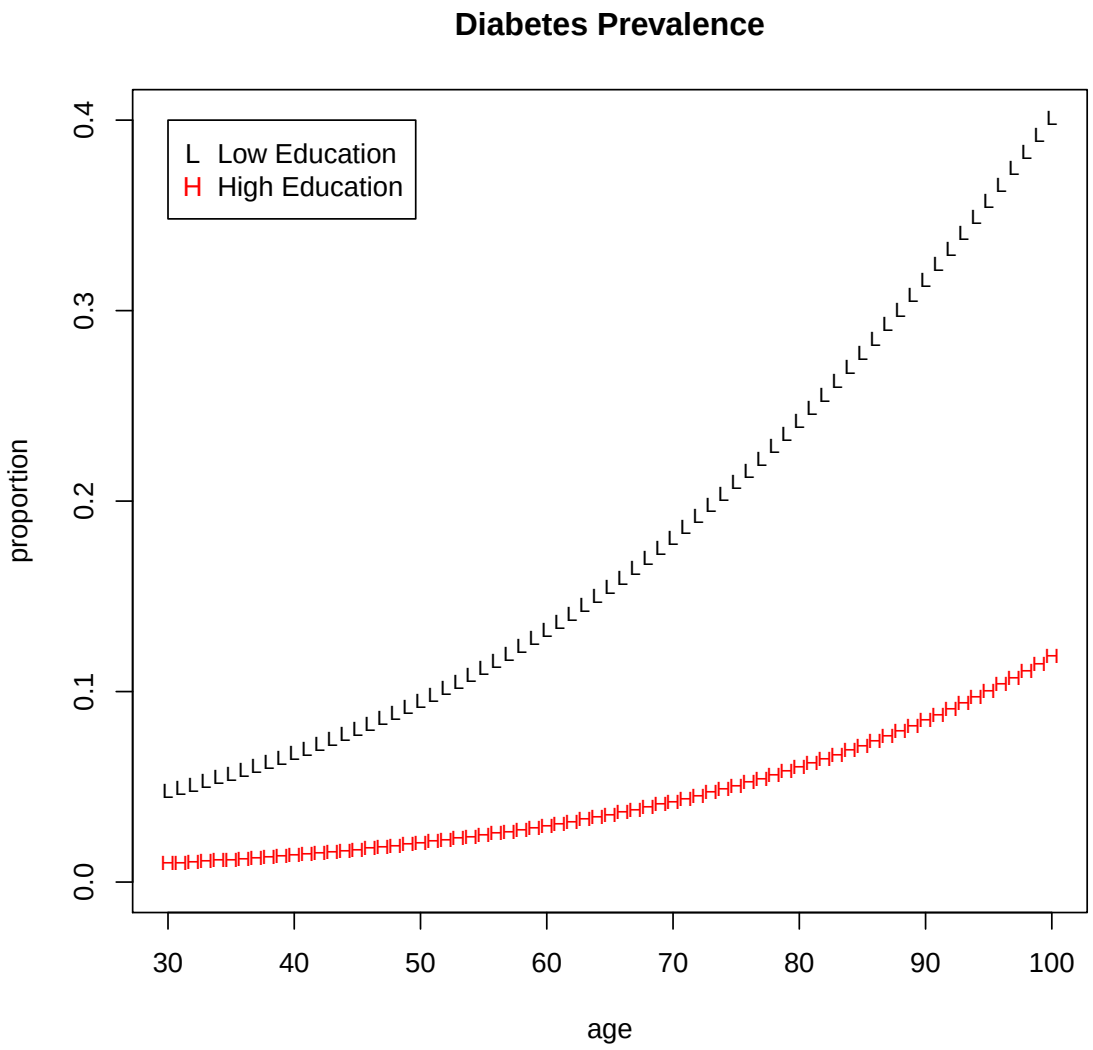


Figure 4: Age−specific probability of having diabetes, by education group, used in sir

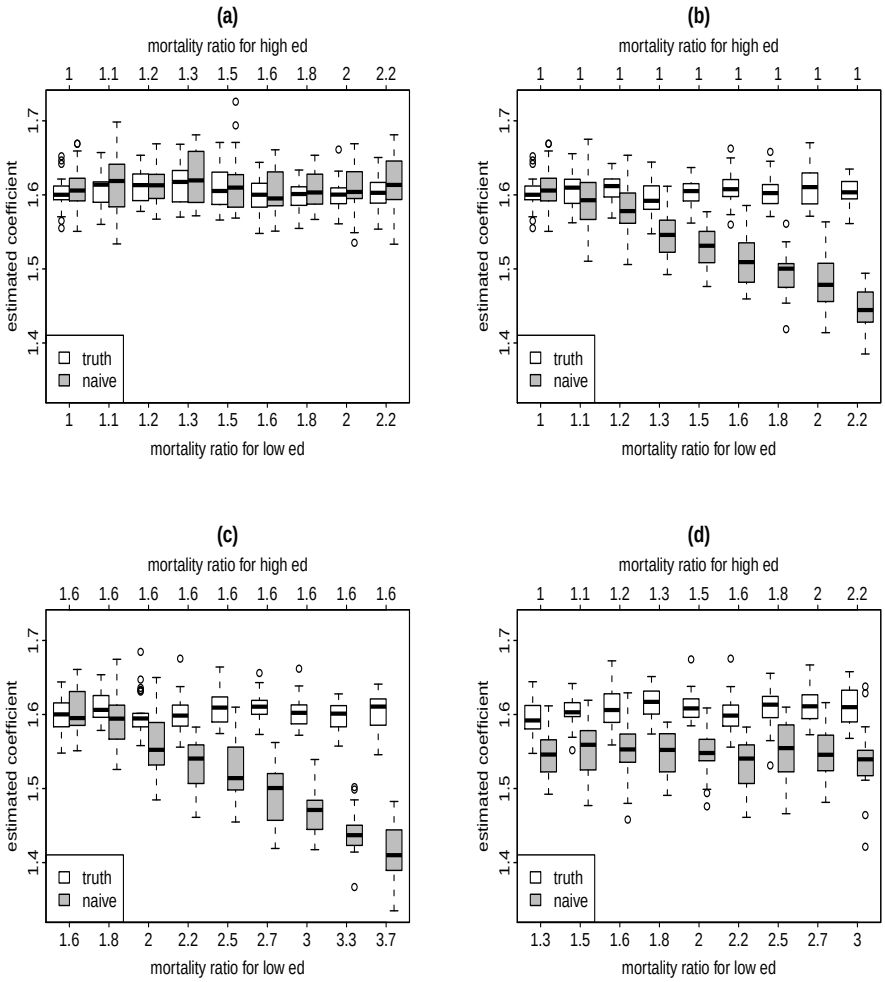


Figure 5: Estimated coefficients for the effect of education on the log odds of having diabetes (conditioning on age) in the presence of differential mortality.

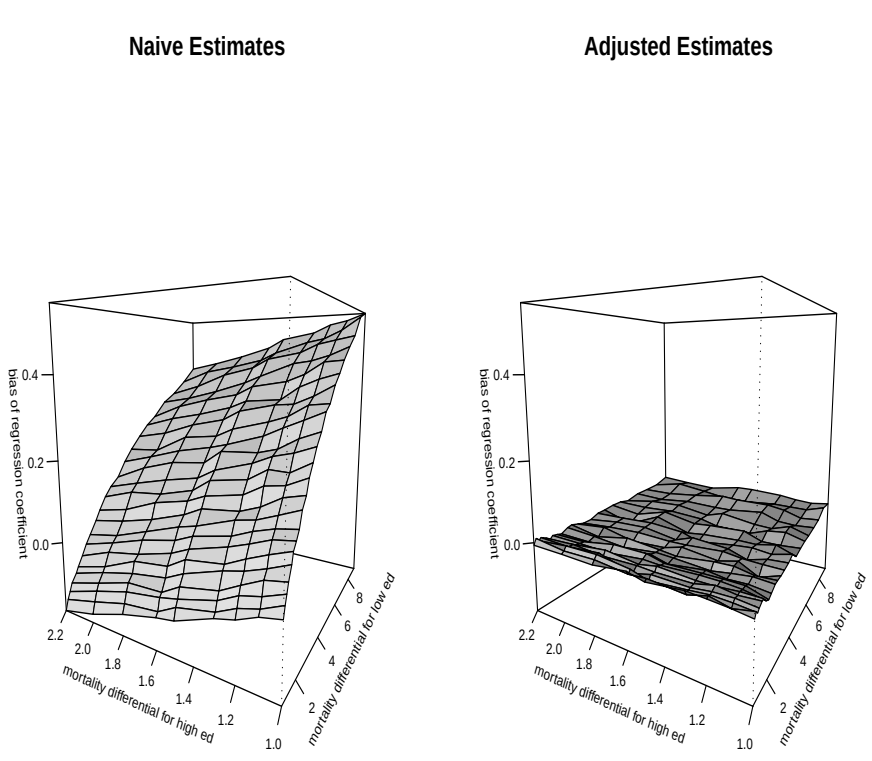


Figure 6: Bias in the estimated coefficients for the effect of education on the log odds of having diabetes (conditioning on age) for the naive and adjusted approaches.

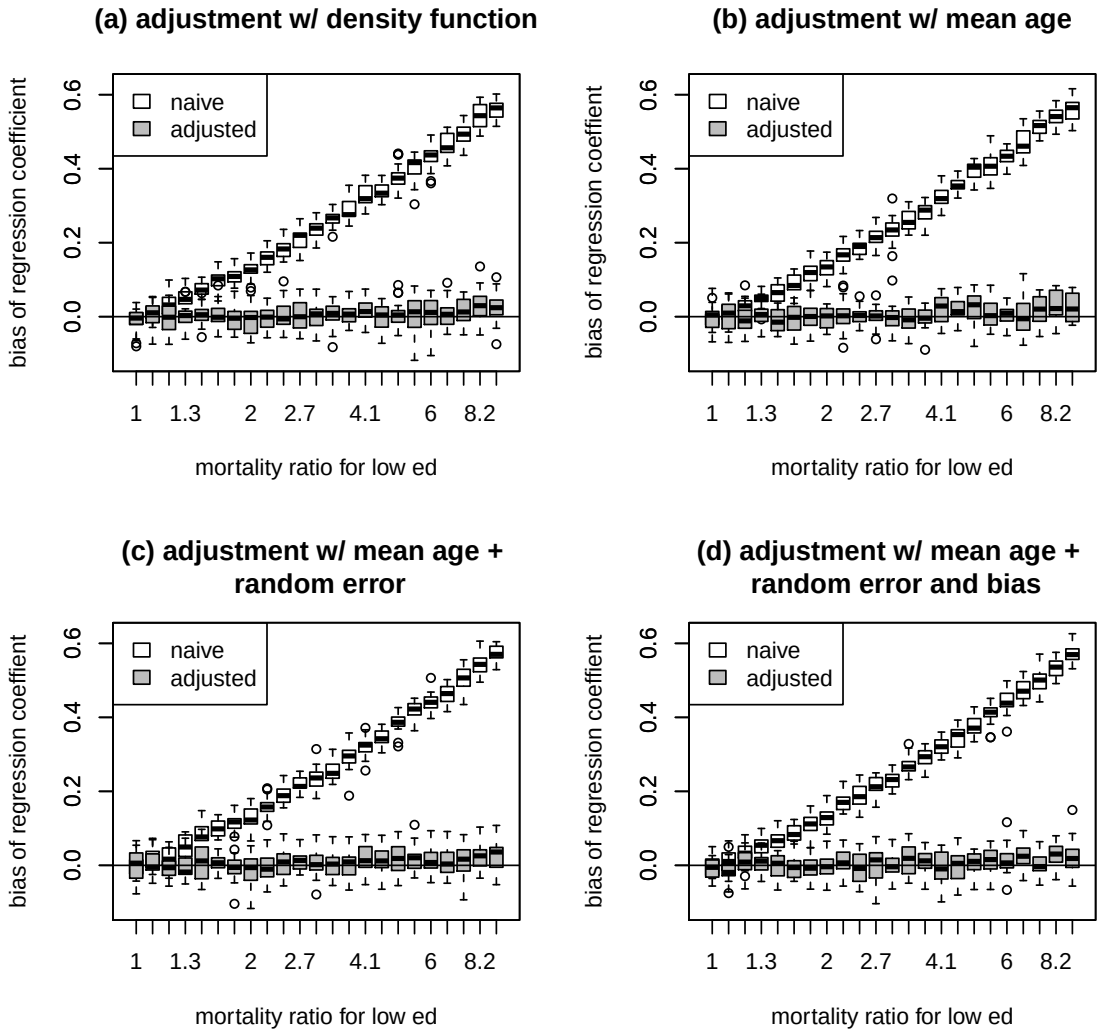


Figure 7: Bias in the estimated coefficients for the effect of education on the log odds of having versions of the adjustment.

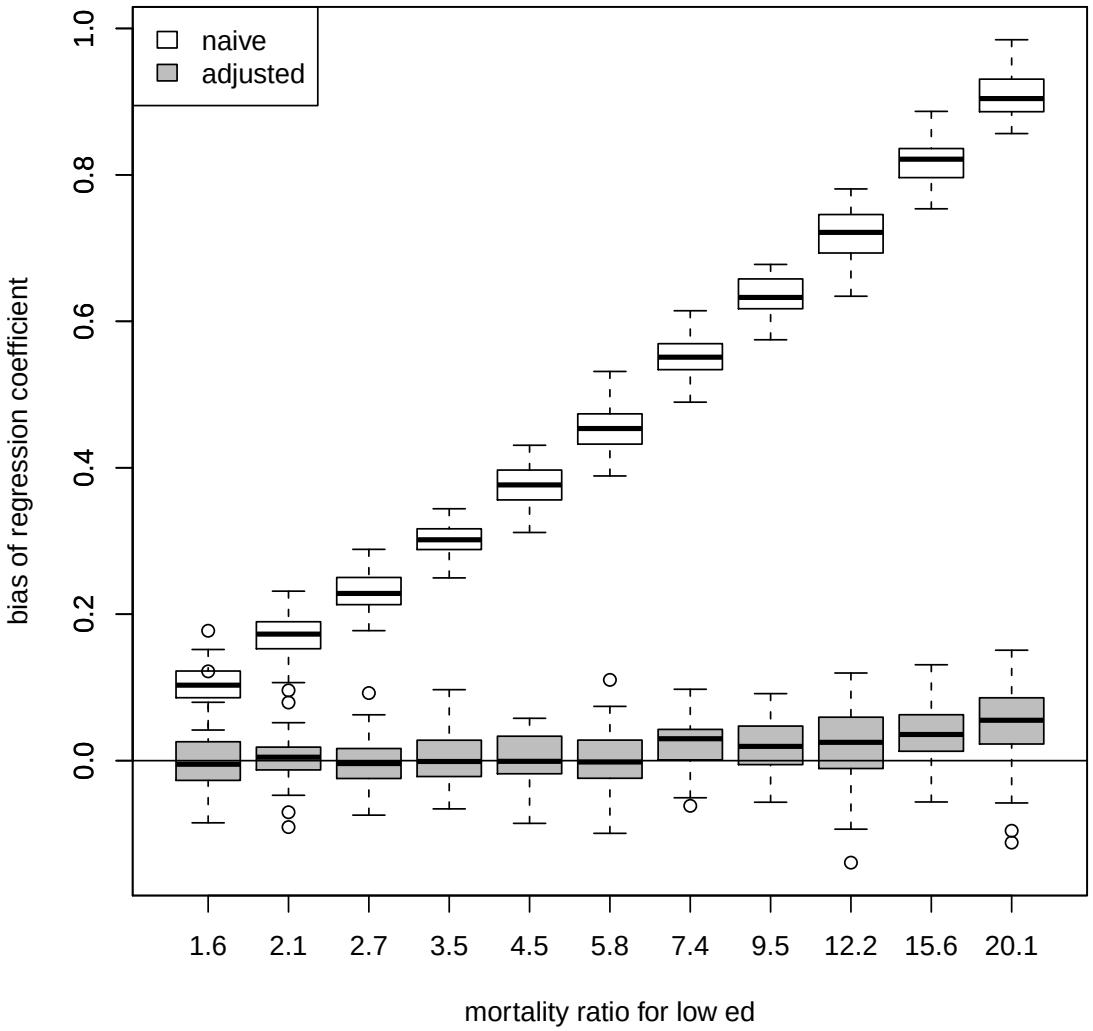


Figure 8: Bias in the estimated coefficients for the effect of education on the log odds in the presence of different distributions of waiting time to diabetes for the low and hi