Center for Demography and Ecology

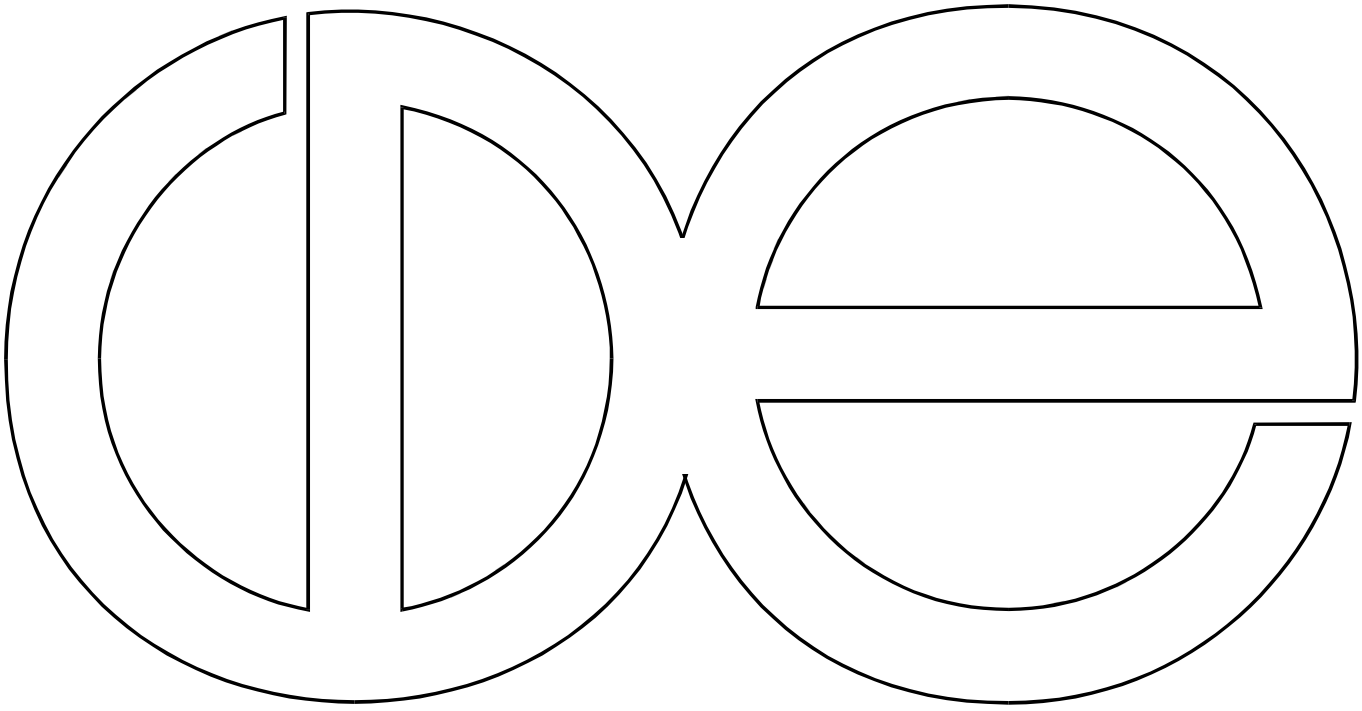
University of Wisconsin-Madison

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

Alberto Palloni

Jason R. Thomas

CDE Working Paper No. 2011-09



ESTIMATION OF HEALTH STATUS INEQUALITIES FROM PREVALENCE DATA: A RISKY BUSINESS

Alberto Palloni
&
Jason R. Thomas

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

Abstract

Assessing the impact that socioeconomic determinants have on the prevalence of certain chronic conditions reported by respondents in population surveys must confront two problems. The first is that the self-reports could be in error (false positives and false negatives). The second is that those reporting are a selected sample of those who ever experience the problem, the selection being heavily influenced by excess mortality due to the condition being reported. In this paper, we use a combination of empirical data and microsimulation to (a) assess the magnitude of the biases due to the selection problem, and (b) suggest adjustment procedures that correct for biases.

INTRODUCTION

Accurate inferences about incidence of phenomena are generally made from data collection plans that follow observations over time and allow precise measurement of the timing of occurrence of relevant events. However, longitudinal designs are expensive enterprises and sometimes researchers replace them by single wave cross-sectional surveys with retrospective recall. For example, a significant number of phenomena such as onset of illnesses, recovery from treatment, menopause, weaning, leaving home, first marriage and the like rely on information collected retrospectively in cross-sectional surveys. But retrospective recall of events and their timing is oftentimes inaccurate and statistical inferences from this information stand on shaky ground.

An alternative to retrospective recalls is current status data, that is, information about the occurrence of a relevant event prior to a time marker, such as the date of a survey. This information is less sensitive to recall problems and can be retrieved easily in conventional interviews. Under some conditions examined below this information and associated statistical tools are a good basis for inferences about the underlying incidence of a phenomenon and for the identification of the determinants of its intensity and duration profile. The information is sometimes aggregated and represented as prevalence data, namely, the fraction of the observations that experiences the event by a time marker. 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 single females at age x is used to identify characteristics of the timing of first marriage (Hajnal, 1953). Or the age-specific proportion of a population who has been diagnosed with diabetes yields useful insights about processes that drive the incidence of diabetes.

Current status data have an important drawback: they rest on the assumption that attrition of individuals who experience the event of interest is the same as the attrition of those who do not. This assumption is violated when the event under study (illness or disability, for example) is associated with at least one source of attrition (e.g. mortality). For example, information about current breastfeeding cannot be elicited from mothers whose most recently born child died before the survey. But these children are more likely to have breastfed for shorter periods of time or not breastfed at all. Similarly, members of a cohort who contract diabetes and then die before the survey as a result of complicating factors cannot provide information about their diabetes status. But their mortality risks were higher than average because of the presence of diabetes. Although this weakness of current status data is well-known, it is conventionally trivialized, dismissed or altogether ignored in empirical applications. In this paper we assess the magnitude of biases when the assumption of homogeneous risks, e.g. identical pre-survey attrition among those who do and those who do not experience the event of interest, is violated, and we develop an adjustment procedure that corrects the bias. The paper is organized as follows: Section 2 reviews an example of the application of current status techniques in population studies. Section 3 identifies the linkages between current status information and the underlying incidence function in a non-formal manner. In Section 4 we introduce population heterogeneity, identify inferential problems when there is differential pre-survey attrition, and propose an adjustment using

maximum likelihood procedures. Section 5 evaluates these procedures with Monte Carlo simulations. Section 6 applies the adjustment procedure to a concrete case, and Section 7 summarizes the results and concludes.

TWO EXAMPLES: THE STUDY OF MARRIAGE AND DISABILITY

We start with two well-known examples. The first is taken from influential work on the history of marriage. The second is drawn from recent controversies about trends in old-age disability in the US.

Timing of Marriage and Proportions Single

Almost sixty years ago now John Hajnal proposed the use of the Singulate Mean Age at Marriage, better known as SMAM, to estimate the mean age at marriage (Hajnal, 1953).[1] It was the first demographic application of current status techniques. SMAM has been widely used to assess the timing of first marriage from census information on the age-specific proportion single. It makes use of the fact that, under some conditions, the age-specific prevalence of singlehood is a good indicator of the (single decrement) probability of remaining single during the interval elapsed between the age at which the population begins to marry and the age of the individual at the time of a census or survey. The first condition is that the incidence of marriage be time invariant (stationarity). The second is that the risks of attrition, mostly induced by mortality or migration, are identical among single and married people (risk homogeneity).[2] A very large literature and influential theories on historical demography are based on inferences based on SMAM or, alternatively, on the proportion single in some target age groups, usually above age 45. In his landmark piece on the so-called Western European marriage pattern, Hajnal made extensive use of these indicators (Hajnal, 1965) to characterize two different continental marriage regimes, a distinction which exerted a great deal of influence on subsequent research on fertility and family formation.

While most analysts are aware of the need to invoke the assumptions identified above, we know of no study that evaluates the potential pitfalls when the assumptions are inappropriate. To begin with, inferences from SMAM can only be correct when there are no time trends. If there are, then the assumption of stationarity is violated and current status proce-

[1] When proportions single are in one-year age groups, the standard expression for SMAM is

$$\text{SMAM}=15+(\sum_{x=15}^{x=49} S_x - 35 * S_{50})/(1 - S_{50})$$

where S_x is the proportion single in the age group (x,x+1). The expression assumes that there is no marriage before age 15 or after age 50.

[2] In what follows risk homogeneity refers to a situation when the risk of attriting before (and hence not being observed by) the time of a census or survey is independent of the event being studied. Conversely, risk heterogeneity is a situation where pre-census (survey) attrition occurs differentially among those who do and those who do not experience the event of interest.

dures should be avoided. By the same token, it is known that, in violation of the assumption of risk homogeneity, mortality risks experienced by single individuals are higher than those experienced by married individuals (Livi Bacci, 1985; Kisker and Goldman, 1987; Hu and Goldman, 1990). If all processes are stationary and mortality risks of single individuals are higher than those of the married population, the observed proportions single will be too small and SMAM will be too large (see definition of SMAM in footnote 1). Simulations with a wide variety of marriage and mortality patterns indicate that sensitivity of SMAM to mortality differentials is not trivial (see Appendix for details on the simulations): a 1% mortality differential can produce proportionate errors in SMAM that are as small as .1% and as large as .8%, depending on the magnitude of the proportion that eventually marries. If the probability of ever marrying hovers around .75-.85, as was the case in Northern and Western Europe in the middle of the XIXth century, a mortality differential equivalent to 10% will bias SMAM upwards by approximately 7%. Hence if the true mean age at marriage is 28 and the probability of ever marrying is .80, the estimate from SMAM could be as high as 31; and if mortality differentials are 20%, the estimate from SMAM would be calculated to be 34![3]

Worse yet, if mortality differentials between single and married individuals decline over time, the observed SMAM falls, producing the appearance of a decline in the mean age at marriage in the absence of any trend. If there is a decline in the mean age at first marriage, as in Northern and Western Europe after 1850, and simultaneously there is a reduction in mortality differentials, the observed trajectory of SMAM will exaggerate the magnitude of the rate of decline in the age at first marriage. Errors of similar magnitude affect inferences about regional or national patterns. For example, the divide between the so-called Western and Eastern marriage pattern based on the observation of higher SMAM and higher proportion single in Western and Northern Europe, depends on the assumption of identical mortality differentials in both regions.

Errors of higher magnitude affect the proportion single at some upper limit, say 50 or 55, beyond which first marriage is negligible. When mortality among singles is higher than overall mortality, the observed proportion single is an underestimate of the true probability of being single. The simulations referred to above indicate that a mortality differential of only 10% translates into a downward bias associated with the proportion single equivalent to 15% in the age interval 45-54.

The key but unanswerable question is about the actual magnitude of the mortality differentials between single and married individuals. Patterns in modern societies (Hu and Goldman, 1990; Livi Bacci, 1985) suggest that at the relatively low levels of mortality prevailing the differentials are not trivial. It is unlikely that the more severe mortality regimes of the past induced lower differentials between the married and single and if so, conventional SMAM-based inferences about historical trends and regional contrasts rest on incorrect information.

Trends in Disability in the US

[3]Results of simulated values of SMAM under different conditions are available on request.

Recent literature suggests that old age (65+) disability in the US has been decreasing (Freedman et al., 2004; Manton et al., 2006; Schoeni et al., 2008). The exact figures are hard to pin down as they are sensitive to the age range being investigated, the measure of disability, the data sources, and adjustments for sampling errors and population representation. Freedman et al. (2004) suggest that a measure of prevalence of disability based on needing help with one or more Activities of Daily Living (ADL) declined from 12% in 1994-5 to about 11% in 1999-2000 (yearly rate of decline of about .017) depending on the data source. Using a different indicator of disability, Manton et al. (2006) estimate that age-standardized prevalence of disability declined from 23.2% to 19.0% between 1994 and 2004-5 (a yearly rate of decline of .018).[4]

Contrasts between social groups are based on similar statistics. For example, Manton and Gu (2001) show that in the age group 65-74 whites' disability prevalence among those with 9-12 years of schooling decreased from 10.3% to 9.0% during the period 1994-1999. The figures for a comparable group of blacks are 21.6% and 21.5% respectively. According to these estimates, blacks experience about twice as much disability as non-blacks and virtually no improvements over time.

Apart from variability stemming from other sources of errors or uncertainty, there is the problem that what is being compared is a current status measure of disability. If there are mortality differentials between the disabled and non-disabled populations then the prevalence measure will not reflect the single decrement probability of disability. Using data from the Longitudinal Study of Aging, Crimmins and colleagues (1997) show that the relative mortality risk among the disabled aged 76-96 during the period 1986-1990 was several times higher than among the non-disabled.[5] If disabled individuals indeed experience higher levels of mortality, the prevalence of disability will understate the true probability of becoming disabled.

It may well be that these errors are irrelevant as researchers are only interested in assessing the load of disability and its cost, not on making inferences about the underlying process reflected in the hazard of disability. But for those interested in testing theories about compression of morbidity, for example, empirical estimates based on current status measures should not be the central quantities. Furthermore, statements about group disability differentials or time trends in the probabilities of becoming disabled will be off, regardless of what the goal of the researcher may be.

Assuming a true level of disability of about 12.75% in the age group 65+ and variable mortality differentials between the disabled and the non-disabled, we estimate that the average elasticity of the proportionate error in the prevalence of disability relative to proportionate differentials in life expectancy at age 65 is about .30. Thus, if mortality among the disabled has been increasing relative to mortality among the non-disabled at an annual rate of say .20 years of life expectancy per year in an interval of five years, the observed prevalence of disability in the age group will be about 1.8% lower than the true probability of becoming disabled. Small as this quantity may seem, it is enough to account for some

[4]The measure was based on impairment in one or more ADLs or instrumental activities of daily living (IADL) for 90 days or more, or use of an institutional residence with medical services available 24 hours a day (e.g. nursing home, assisted-living residence).

[5]There is some evidence of an increase (decrease) in the risk of mortality among the disabled (nondisabled) after the period 1984-86, but the differences over time were not statistically significant.

of the declining trends documented in the literature. The same applies to comparisons of disability between blacks and non-blacks: the observed difference is likely to be too small to represent underlying differences in the rates of becoming disabled as excess disability-related mortality among blacks is likely to be larger than among non-blacks.

THE ALGEBRA OF CURRENT STATUS DATA

Suppose we are interested in the occurrence of event e. Let i index individuals characterized by a waiting (latent) time or duration, d_{ei}, defined as the elapsed time between the calendar time of onset of exposure, t_{oi}, and the calendar time of the occurrence of the event, t_{ei}, as well as by a probability P_e that the event will ever be experienced. We assume that individuals are observed at an exact date, t_s, the date of a survey which is identical for all individuals and independent of e. The survey provides enough information to define an indicator variable $\eta_i = I_i(t_{oi} < t_{ei} \leq t_s) = 1$ if the event takes place before t_s and 0 otherwise. Occasionally, but not always, surveys also contain retrospective questions to elicit the timing of the event, t_{ei}. In general, however, this information is not collected or is unreliable. If the event took place before the survey, 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 have right-censored data. If individuals provide 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 the passage of time so that we can translate the above time and duration indicators into an age metric. Let x_{oi} 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 the event (if this took place) so that duration is $d_{ei} = x_{ei} - x_{oi}$. When d_{ei} is known it is usually subject to considerable noise and in what follows we 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 time invariant trait or independent variable, Z, which the investigator believes exerts influence on the occurrence and timing of event e but not on the timing of the survey, t_s. The observable (aggregate) 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 $\overline{N}(x, t_s, Z = 0)$.

It is possible that at time t_s we will observe all individuals who experienced e because there is a set of censoring events $\{c = 1...k\}$ each characterized by a duration d_{ci} for individual i such that if $d_{ci} < (t_s - t_{ei})$ we will have *no information whatsoever on the individuals.*[6]

[6]An event c censors observation of individual i when $t_{ci} < t_s$. Left censoring is problematic in current status data when $t_{ei} < t_{ci} < t_s$, censoring precludes collection of information on the event of interest, and c and e are not independent. It should be noted that the problem we face is one where there is no information at all on individuals who are left censored by alternative events. Note that this is quite different from the standard (Heckman-type) selection problem where some individuals are only partially observed.

For example, in a cross-sectional survey of older people, the researcher may have information on diabetes status for all those who survived to age x and no information at all on individuals who died (with or without diabetes). 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 one representing events leading to unobserved individuals at time t_s. The notation in this figure makes explicit an important assumption we use throughout the paper, namely, that all processes are stationary (do not depend on calendar time).

Let $\delta(y, Z = 1)$ be the instantaneous risk at age y of event e for individuals with trait $Z = 1$, $\upsilon(y, Z = 1)$ the instantaneous risk at exact age y of alternative censoring events among those with trait $Z = 1$ who experience e and, finally, $\mu(y, Z = 1)$ the instantaneous risk at age y of alternative censoring events among those with trait $Z = 1$ who do not experience e. The investigator is interested in the function $\delta(y)$ referred to as the "incidence" function, the "risk" function and the "hazard" function of event e and on the effects that Z may have on it. To simplify exposition assume that there is only one alternative censoring event, say mortality, and that the risk of this event is dependent only on age (not on duration since the occurrence of e). All results differ only slightly if the function $\upsilon(.)$ is duration dependent.

The process represented in Figure 1 has been well-studied by Keiding (1991; 2006) and by authors interested in statistical inference from current status data (Diamond and McDonald, 1991; Sun and Kalbfleish, 1993; Keiding et al., 1989; Keiding et al., 1996). But detailed attention to the problem generated by the existence of censoring has only recently been formally investigated (Jewell and Van der Laan, 2004).[7] This paper rests on some of these developments and proposes a tractable solution for empirical estimation. In what follows we introduce the basic algebra of current status data and derive expressions in the case of homogeneous and non-homogeneous risks. We first deal with the case when the risk of event e is the the same for all individuals and then when there is a binary trait or independent variable Z that affects the incidence of e.

Case 1: Homogeneity of Risks[8]

Assume there are no mortality differentials between those who experience e and those who do not, e.g. $\upsilon(r) = \mu(r)$, 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

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

[7] Lin et al. (1998) study the case of differential mortality and propose a model for current status data collected in an experimental setting in which all subjects are observed (and the monitoring time depends on the event of interest). We consider a different situation in which current status data are randomly sampled from a population, and differential mortality is more likely to prevent population members who experience the event of interest, relative to those who do not, from surviving to (and thus being observed at) the time of the survey (for individuals of a given age).

[8] What follows is a tight summary of materials that has already been the subject of excellent reviews (Keiding, 1991; 2006).

where $\Phi(x) = \exp(-\int_0^x \mu(r)dr)$ is the (single decrement) probability of surviving to age x and $\Lambda(x) = \exp(-\int_0^y \delta(r)\ dr)$ is the (single decrement) probability of avoiding event e. If, as happens with many conditions with adult onset, the process starts at an age $x_o > 0$, the above expressions hold but with an origin shift. Inferences about the process(es) leading to the occurrence of e focus on the function $\delta(r)$ or any of the quantities defined by it, particularly the integrated hazard, namely, $(\int_0^y \delta(r)\ dr)$. The observed data can be used to make inferences about $\delta(r)$ and the set of parameters on which it depends. Our purpose is to show that such inferences can only be made under restrictive assumptions regarding alternative censoring events.

If the number of entrances at origin x years before the survey 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 experiencing e and surviving to age x at time t_s is

$$\Omega(x) = \int_0^x \exp(-\int_0^y (\mu(r) + \delta(r))dr) * \delta(y) * \exp(-\int_y^x \mu(t)dt\ )\ dy$$

or

$$\Omega(x) = \exp(-\int_0^x \mu(r)dr) * \int_0^x \delta(r) * (\exp(-\int_0^y \delta(t)dt)dy = \Phi(x) * (1 - \Lambda(x)).$$

The expected number of surviving individuals aged x who experience e before t_s is

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

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

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

and, conversely, the observed proportion who have not experienced e, $\overline{p}(x, t_s)$, 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 (Keiding, 1991).[9]

Case II: Heterogeneity of Risks

Suppose that $\mu(r) \neq \upsilon(r)$ and $\mu(r)$ is a baseline mortality risk that applies to the

[9]Thus the conditions under which observed prevalence is an unbiased estimate of $\Lambda(x)$ are thus very general. Even under irregular birth streams do the equalities hold. They only break down when the population is open to migration and migration flows are associated with the risks of events involved.

population that does not experience the event. The expression for the function $\Omega(x)$ becomes

$$\Omega(x) \;=\; \int_0^x \exp(-\int_0^y(\mu(r)+\delta(r))dr) * \delta(y) * \exp(-\int_y^x \upsilon(t)dt)dy$$

Further simplification is possible if 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 mortality risk differential. After simplification, the expression for $\Omega(x)$ becomes

$$\begin{aligned}
\Omega(x) &= \int_0^x \exp(-\int_0^y(\mu(r)+\delta(r))dr) * \delta(y) * \exp(-\int_y^x \upsilon(t)dt)dy \\
&= \exp(-\int_0^x \mu(r)dr) \int_0^x \exp(-\int_0^y \delta(r)dr) * \delta(y) * \exp(-\int_y^x(\upsilon(t)-\mu(t))dt)dy \\
&= \Phi(x)(1-\Lambda(x)) \int_0^x \tfrac{\exp(-\int_0^y \delta(r)dr)*\delta(y)}{(1-\Lambda(x))} \exp(-\int_y^x(g(\theta)\mu(t)-\mu(t))dt)dy \\
&= \Phi(x)(1-\Lambda(x)) \int_0^x \varphi(y) * \exp\{-(g(\theta)-1)\int_y^x \mu(t)dt\}dy
\end{aligned}$$

where $\varphi(y)$ is the conditional density of the waiting times of event e at age y given that the event occurs by age $x > y$. Using the mean value theorem we re-express $\Omega(x)$ as:[10]

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

where

$$\alpha(x, \widetilde{y}_x) = \exp(-(g(\theta)-1)\int_{\widetilde{y}_x}^x \mu(t)dt),\ \widetilde{y}_x\ ,\ (0 < \widetilde{y}_x < x)$$

is an implicit function of $\varphi(.)$, θ, and $\mu(.)$. The observed proportion who experience event e by age x is

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

whereas the observed proportion who survived without experiencing E is

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

When $g(\theta) = 1$ then $\alpha(x, \widetilde{y}_x) = 1$ and we are back to a situation of (mortality) risk homogeneity. When $g(\theta) \neq 1$, individuals who experience e are exposed to different mortality risks than the general population and the value of $\alpha(x, \widetilde{y}_x)$ can be smaller or larger than 1. As a consequence, the observed proportions who experience event e do not provide enough information to retrieve estimates of the hazard or integrated hazard of 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 experiencing the event before age x. For

[10] Under standard regularity conditions, the expression $\int_0^x \varphi(y) * \exp\left(-(g(\theta)-1)\int_y^x \mu(t)dt\right)dy$ can be approximated as $\exp\left(-(g(\theta)-1)\int_{\widetilde{y}}^x \mu(t)dt\right) * \int_0^x \varphi(y)dy$; but since $\int_0^x \varphi(y)dy = 1$, it reduces to $\exp\left(-(g(\theta)-1)\int_{\widetilde{y}}^x \mu(t)dt\right) = \alpha(x, \widetilde{y}_x)$.

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 $\Lambda(x_{\max}) = .50$. Since $\alpha(x, \widetilde{y}_x)$ decreases with age when $g(\theta) > 1$, the error in the observed prevalence rates will increase at least up to x_{max}. If $\Lambda(\infty) < .50$ and $g(\theta) > 1$ then the observer will be fooled into believing that *older individuals are less likely to experience e*.

Figure 2 displays the magnitude of proportionate errors in $\Lambda(x)$ in four different cases representing mortality differentials between diabetics and non-diabetics.[11] The age pattern of prevalences contains downward biases that worsen with age. To a naive observer this pattern might suggest that the incidence of e is less intense for older cohorts.

PRESENCE OF COVARIATES

We now assume that a trait Z exerts an effect on the occurrence of e; the goal is to make inferences about the effect of Z on the occurrence of 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 (mortality) risk homogeneity the observed prevalence in each subgroup is sufficient to obtain unbiased estimators of the true single decrement survival probabilities of experiencing e. In such case a contrast of prevalence rates across subgroups yields an unbiased estimator of the effect of trait Z on the risk of contracting the condition. This will not occur under a regime of (mortality) risk heterogeneity in any of the groups.

Estimation of Effects

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 subgroups should be related as follows:

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

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 λ. If the assumption of risk homogeneity

[11] The values of $\widetilde{y}$ were set to be 45, 50, and 55 at ages 60, 65 and 70, respectively. At age 70 and above we set it to be 60. 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 attains a maximum value of .581 at 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 $\alpha_1$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_1 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.

is accurate, an estimate of λ is, in principle at least,[12] retrievable from observed prevalence data. In fact, under such conditions the ratio of prevalence rates in the two subgroups at age x is given by:

$$\begin{aligned} O(x) &= \overline{p}(x, Z=1)/\overline{p}(x, Z=0) = \Lambda(x, Z=1)/\Lambda(x, Z=0) \\ &= \Lambda(x, Z=0)^{\exp(\lambda)-1}. \end{aligned}$$

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

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

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) 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 they can be improperly signed.[13]

In summary, inferences about the occurrence of an event and/or of the magnitude of covariate effects will contain biases when there is (mortality) risk heterogeneity. The approximations illustrated in Figures 2 and 3 suggest that ignoring mortality differentials leads to underestimation of cumulative incidence if e is associated with higher mortality. In general the bias will get worse as duration from the time of onset e increases. In many observed cases of interest (e.g. diabetes) age-specific prevalence rates tend to bend downwards with age, as Figure 2 suggests they would. This does not necessarily mean that incidence among older cohorts is lower. Similarly, data on prevalence of diabetes by levels of education shows a notorious regularity: the curve for the least educated converges and sometimes drops below the prevalence rates among those with higher levels of education. As shown by Figure 3 this could simply be a result of differences in mortality of diabetics and non-diabetics in one or both subpopulations. While in some cases this may reflect true incidence differentials, in other cases it could be the combined result of biases and actual incidence.

A Likelihood Approach

The likelihood of a sample of observations of current status when there is no mortality differential is

[12] Estimation of λ may not always be possible even under these simplifying conditions. But homogeneity of risks drastically simplifies the task.

[13] The values of Λ and α and the four combinations created with them are the same as for Figure 2.

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} \{\exp(-I^m(x_i)(1-\exp(-I^e(x_i)))\}^{Y_i}\{\exp(-(I^m(x_i)+I^e(x_i)))\}^{(1-Y_i)} \\ &= \prod_{i=1}^{N} \exp(-I^m(x_i)) * \{(1-\exp(-I^e(x_i)))\}^{Y_i}\{\exp(-(I^e(x_i)))\}^{(1-Y_i)}\end{aligned}$$

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; Y_i =1 if an individual i experienced the event before or at time t_s and 0 if she did not. Upon conditioning on the probability of surviving up to t_s (dividing by the probability of surviving up to t_s) the term $\exp(-I^m(x_i))$ drops out of the expression.[14] In conventional current status models the researcher specifies the nature of $\delta(y)$ (and therefore p_{x_i}), assumes dependency on a vector of covariates and associated parameters γ, and obtains estimates via standard likelihood procedures. But this is meaningful only because the baseline mortality function carries no information about the incidence of e.[15] When there is differential mortality between those who experience and those who do not experience e the likelihood is given by:

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} \exp(-I^m(x_i))\{\exp(-I^e(x_i))\}^{(1-Y_i)} * \\ &\quad \{(1-\exp(-I^e(x_i)))\}^{Y_i}\{\textstyle\int_0^{x_i} f^e_{x_i}(y)\exp(-(g(\theta)-1)I^m(y,x_i))dy\}^{Y_i}\end{aligned}$$

where $f^e_{x_i}(y)$ is the conditional density $(\delta(y)/\int_o^{x_i}\delta(y)dy)$, $I^m(y,x_i)$ is the integrated hazard from age y to x_i for those who do not experience the event, and $g(\theta)$ is the measure of excess mortality among those who do (see previous section on the heterogeneity of risks for more details). The expression can be rewritten as

$$\begin{aligned}\mathcal{L} &\propto \prod_{i=1}^{N} \exp\{(-I^e(x_i))\}^{(1-Y_i)}\{(1-\exp(-I^e(x_i))\}^{Y_i} * \\ &\quad \{\textstyle\int_0^{x_i} f^e_{x_i}(y)\exp(-(g(\theta)-1)I^m(y,x_i))dy\}^{Y_i}.\end{aligned}$$

The inner integral in the likelihood is the (conditional) expected value of the function $\exp(-(g(\theta)-1)I^m(y,x_i))$ which, using the delta method, can be approximated as:

[14] This simplification rests on the assumption that the timing of the survey, t_s, is independent of the events of interest or the same for all individuals.

[15] The likelihood advocated by Diamond and MacDonald (1991) is one where mortality (or any other type of alternative attrition) is treated as ignorable.

$$\begin{aligned}\int_0^{x_i} f^e_{x_i}(y)\exp(-(g(\theta)-1)I^m(y,x_i))dy &\cong \exp(-E_{x_i}((g(\theta)-1)I^m(y,x_i)))\\ &\cong \exp(-(g(\theta)-1)I^m(E_{x_i}(y,x_i)))\\ &= \exp(-(g(\theta)-1)I^m(\widetilde{y}_{x_i},x_i))\\ &= \exp(\varphi(\theta;\widetilde{y}_{x_i}))\end{aligned}$$

where E_{x_i} stands for 'conditional expectation at age x_i with respect to the density $f^e_{x_i}(y)$' and the quantity $\widetilde{y}_{x_i}$ is the mean age at which the event is experienced (conditional on experiencing it before x_i).[16] The likelihood simplifies to

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

and no factorization of risks is possible. Thus, in the presence of heterogeneous risks the conventional strategy of current status data is no longer feasible.

Conditioning on survival to t_s we get the conditional likelihood of the sample:

$$\begin{aligned}\mathcal{L} \propto \prod_{i=1}^{N}&\{(1-p_{x_i})/((1-p_{x_i})+p_{x_i}\exp(\varphi(\theta;\widetilde{y}_{x_i})))\}^{(1-Y_i)}*\\ &\{p_{x_i}\exp(\varphi(\theta;\widetilde{y}_{x_i}))/((1-p_{x_i})+p_{x_i}\exp(\varphi(\theta;\widetilde{y}_{x_i})))\}^{Y_i}.\end{aligned}$$

This expression can be generalized to cases where the researcher is interested in the effects of covariates contained in a vector Z and a vector of associated parameters γ. One way to make $\mathcal{L}$ tractable is to assume that the log of the waiting times is 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 (including $ln(x_i)$) and β is a vector of effects. Replacing this in the conditional likelihood we obtain after simplification

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

[16] This expression for $\varphi(\theta;\widetilde{y}_{x_i})$ assumes that the mortality of those who experience the event is proportional to the mortality of those who do not. In more general cases we will express the mortality of both groups as proportional to some common standard or baseline. In this case the expression for $\varphi(\theta;\widetilde{y}_{x_i})$ should be $(g(\theta)-1)I^s(\widetilde{y}_{x_i},x_i)$ where $I^s(\widetilde{y}_{x_i},x_i)$ is a standard integrated hazard.

or the likelihood of a conventional logistic model with the set of covariates expanded to include $\varphi(\theta; \widetilde{y}_{x_i}, x_i)$.[17]

What Should $\varphi(\theta; \widetilde{y}_{x_i})$ Be?

The quantity $\varphi(\theta; \widetilde{y}_{x_i})$ stands for the difference of integrated hazards between those who experience and those who do not experience the event. The lower limit of the integrated hazard is the *expected* age at which e occurs to those who are aged x_i. It is the difference in the *predicted* probabilities of being 'eligible' (alive) for the survey after the occurrence of e 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 one knows (a) the incidence curve and its determinants, (b) the force of mortality $\mu(y)$, and (c) the parameter of excess mortality, θ. First, the incidence curve can be approximated from retrospective (but possibly erroneous) information about the timing of e or from known incidence curves of e in populations similar to the one under study. We show below that the adjustment we propose is largely insensitive to minor variability of the values of $\widetilde{y}_{x_i}$. Second, throughout our discussion $\mu(y)$ refers to the mortality risks among those who do not experience E. This quantity is unlikely to be known with any precision. However, to implement an adjustment procedure, it suffices that we identify a standard (baseline) age pattern of mortality that applies to both those who experience and those who do not experience the event.

Estimation with Population and Risk Heterogeneity

Assume two subgroups defined by a binary variable Z. Assume also that each of them experiences risk heterogeneity that can be parameterized using a unique standard pattern of mortality. This is equivalent to defining $\upsilon(x; Z = 0) = e^{\theta_0}\mu_s(x)$ and $\mu(x; Z = 0) = e^{\vartheta_0}\mu_s(x)$ for the first subgroup ($Z = 0$), and $\upsilon(x; Z = 1) = e^{\theta_1}\mu_s(x)$ and $\mu(x; Z = 1) = e^{\vartheta_1}\mu_s(x)$ for the second ($Z = 1$). Thus the integrated hazard $I^m(\widetilde{y}_{x_i}, x_i)$ can be computed using a unique function $\mu_s(x)$ for all observations. With this parameterization we can define a logistic model including the following independent variables: $ln(x_i)$, a dummy variable Z, the integrated hazard $I^m(\widetilde{y}_{x_i}, x_i)$ and the interaction term $I^m(\widetilde{y}_{x_i}, x_i) * Z$. One can show that the estimated coefficient of Z corresponds to the effect of Z (subgroup) on the incidence of E, the estimated coefficient of $I^m(\widetilde{y}_{x_i}, x_i)$ corresponds to $(e^{\theta_0} - e^{\vartheta_0})$, and the estimated coefficient of the interaction term corresponds to $((e^{\theta_0} - e^{\vartheta_0}) + (e^{\theta_1} - e^{\vartheta_1}))$. Thus, under these assumptions we can always retrieve the effects of Z as well as estimates of the mortality differential between those who experience e and those who do not. However, the parameters that define mortality risks for each subgroup cannot be identified. But even under these conditions, the two extra terms containing the integrated hazard are sufficient to obtain adjusted estimates of the effect of Z. Thus, the parameters of the mortality differentials are ancillary and not of central interest.

[17] If functional forms other than the logistic are deemed appropriate, the same conclusions about biases and inferential difficulties apply and only the functional form of the adjustment factor changes.

In most real situations, we will have no information on the age when e occurred. If so, the function $I^m(\widetilde{y}_{x_i}, x_i)$ will attain the same value among those who do and those who do not experience E. When auxiliary information, even if defective, on the age at which the event took place is available, individuals who experience it can be assigned a value of $I^m(\widetilde{y}_{x_i}, x_i)$ with $\widetilde{y}_{x_i}$ replaced by the observed age. In either case, the precision of the adjustment will depend on the coarseness of age measurement.

Current Status, Unmeasured Heterogeneity, and Sample Selection Bias

The problem formalized above is a member of a more general class of problems characterized by two features: (1) the phenomenon of interest is only partially observed; and (2) whether or not the observation takes place is partly determined by the phenomenon being studied or by characteristics that influence the occurrence of the phenomenon. One of the more well-known members of this class is the so-called *unmeasured heterogeneity problem* (Vaupel et al. 1979; Manton and Stallard 1981; Heckman and Singer 1984; Trussell and Richards 1985; Vaupel and Yashin 1985; Trussell and Rodríguez 1990; Hougaard 2000) which arises in longitudinal data whenever the occurrence of the event of interest is a function of variables that are unmeasured (or ignored). A consequence of this omission is that both inferences about the incidence of the event and effects of covariates are erroneous. This is similar to the problem we study in this paper except for one major difference: in the standard case of unmeasured heterogeneity the researcher possesses time-dependent information on all individuals, including those who cease to be exposed to the event of interest, and thus some adjustments are possible (e.g. Heckman and Singer 1984; Trussell and Richards 1985). These adjustments, however, cannot be implemented for the current status problem we are concerned with because individuals whose current status cannot be assessed at the time of the survey are not observed at all. If the researcher is able to collect *repeated current status* information over time on an initial sample of individuals, then the situation will resemble and indeed converge to the standard unmeasured heterogeneity problem.

Another close kin of the current status problem (as formulated here) is the classic *sample selection problem*, where an outcome of interest is observed only among a subset of sample members who differ systematically (from the rest of the sample) on observed characteristics (Heckman 1979; Berk 1983; Greene 1981; Fligstein and Wolf 1978; Wooldridge 1995; Little 1995). Indeed, ascertaining the current status of a non-random subsample of observations is akin to the sample selection problem. But, here again, the difference is that with the latter, the researcher can implement adjustments using information available for *all individuals*. For the current status problem we are concerned with, no such adjustments are possible since the researcher does not have *any information at all* for relevant individuals who are not observed because the sample is left-truncated.

MONTE CARLO SIMULATION

In this section we evaluate the proposed adjustment using Monte Carlo simulations of

a heterogeneous population consisting of two education groups with risk heterogeneity. We consider a number of scenarios characterized by different levels of risk heterogeneity and in each case we estimate parameters of interest and associated biases with and without the adjustment procedure.

Simulated Populations

Consider the population at some time t_s when members range in age from 31 to 100 years and have been exposed to the risk of both dying and becoming diabetic. We choose to start exposure at age 30 so that the youngest cohort of survivors to time t_s (observed at age 31) has been exposed to both risks for one year whereas the oldest cohort of survivors at time t_s and observed at age 100, has been exposed to both risks for seventy years. We assume that the size of the birth cohorts are unequal and that the initial size of each grows at rates between .001 and .005 per year. With a radix of 1000 this yields a total of about 50,000 individuals in each of the subgroups we define below.[18]

We assume that the log of the waiting time to developing diabetes follows a logistic distribution with a constant variance and a mean that is higher among those who have high education relative to those with low education. In the absence of mortality the prevalence of diabetes at age 100 is expected to be roughly 20% among the high education group and just over 40% among the low education group. The resulting regression of the log odds of being diabetic on the log of age and education yields a (true) coefficient of 1.00 associated with low education. The force of mortality follows a Gompertz function with a level parameter that varies by education group. We simulate a number of scenarios that vary in terms of the magnitude of the mortality differentials between diabetics and non diabetics. Each of the resulting scenarios is simulated 25 times.[19]

Under risk heterogeneity once an individual develops diabetes, the risk of mortality increases relative to those without diabetes. As individuals who contract diabetes are exposed to higher mortality, the observed prevalence of diabetes in the group will increase with age less rapidly than it would in the absence of mortality. The cumulative probability of being diabetic is shown in Figure 4 by age and education. This curve is equivalent to the population prevalence of diabetes in the absence of mortality (labeled "No Mortality"). Figure 4 also displays the observed prevalence of diabetes in the low education group whose members are exposed to diabetic-specific mortality rates that are roughly 70% higher than among non diabetics with low education. This is reflected in departures from the expected values in the absence of mortality. Conversely, no mortality differentials between diabetics and non diabetics are assumed in the high education group. This is reflected in observed prevalence rates that fluctuate randomly around expected probabilities of ever contracting diabetes (with increasing variance as the cohort sizes decrease with age).

[18]We simulate growing birth cohorts simply to mimic real populations that are more likely to be growing than remaining stationary. All results apply if all rates of growth are set equal to zero.

[19] Although we started with a large number of simulations we noted that a small number was enough to produce sufficient Monte Carlo variation. As a consequence we settled on a total of 25.

Estimates, Biases and Adjustments

A common approach to assess the size of education differentials in diabetes is to fit a logistic model to the log odds of being diabetic regressed on a constant, a dummy variable for the low education group, some control variables, and the log of age.[20] In the absence of risk heterogeneity, the estimated coefficient of the education dummy variable will reflect (on average) the difference between the two solid lines shown in Figure 4. If there are mortality differentials in the low education group the estimated effect of the education dummy will reflect the difference between the two sets of symbols plotted in Figure 4. These estimates lead to the erroneous inference that education has no effect on diabetes. The magnitude of the bias depends on the magnitude of the mortality differential in the low education group.

Estimates and biases. Our primary focus is the estimated coefficient for the dummy variable identifying the low education group, and its variation with the magnitude of the mortality differential in this education group.[21] The main results are presented in panel (a) of Figure 5. The axis at the bottom of the plot shows the factor by which mortality among diabetics exceeds that of non-diabetics in the low education group: it increases from 1 to 2. Among the highly educated we assume no differentials (this is indicated by the axis on the top of the plot, which is fixed at a value of 1). When there is no mortality differential in either education group, the estimated coefficients for the (low) education dummy variable center around 1, the true value in the simulation. As the size of the mortality differential among the low education group increases, the values of the estimated coefficients tend to zero, as expected.

Figure 5 also shows results for cases when there are mortality differentials in *both* education groups. As pointed out earlier, if the magnitudes of the mortality differentials are the same, then there will be no bias in the estimated effect. This is shown in panel (b) of Figure 5, where there are only small, random fluctuations in the coefficients for each level of the mortality differentials shown. In panel (c) of Figure 5 the mortality differential is fixed for the high education group, while it increases for the low education group. When the mortality differential is larger in the high education group there is an upward bias, and when the differential is larger in the low education group there is a downward bias. The final panel in this figure displays the bias as the size of the differential in the low education group increases at the same rate as the differential in the high education group.

Adjustments. The adjustment procedure requires that we estimate the logistic regression model including the integrated hazard of mortality for a suitable standard population. The new model should include an intercept, the log of age, a dummy variable for low education, the integrated hazard of mortality, and its interaction with the education dummy. Results of fitting a logistic model to the data are displayed in Figure 6. We show the bias (true minus estimated values) in the coefficient of the dummy for education for different combinations of mortality differentials in the low and high education groups when using the unadjusted and the adjusted models. The values plotted in the figure are the mean bias, with the average

[20] Some researchers prefer to fit probit models to prevalence data and then make inferences about incidence (Smith, 2007). While the biases we illustrate in the simulation only apply to logistic models, their magnitude is unlikely to be very different when other functional forms are used to represent current status information.

[21] Recall that in the simulation this coefficient has a true value equal to 1.0.

taken over the twenty-five simulated data sets.

The unadjusted estimates exhibit the same biases described earlier. Thus, if the differential is the same in each education group, then there is no bias but if the differential is larger in the low education group, then the estimated coefficient for the education dummy variable will have a negative bias. Conversely, if the differential is smaller in the low education group, then the estimated coefficient will be upwardly biased.

The adjusted estimates are obtained after controlling for the baseline (standard) integrated hazard. In all cases the integrated hazard associated with individuals aged x is evaluated using the conditional mean of the age of onset of diabetes in the *true* incidence curve as the lower bound of integration. Figure 6 displays the bias from the adjusted logistic model under various scenarios defined by the magnitude of risk heterogeneity. The average bias from the adjusted model forms a flat plane close to zero and the mean adjusted estimate is within a few percentage points of the true effect. Contrast this to the 30% (negative) bias of the unadjusted estimate when diabetes-specific mortality in the low education group is four times higher than non-diabetic-specific mortality.

These simulation results suggest that the proposed adjustment procedure is effective under the conditions used to simulate the data. However, the adjusted estimates are obtained using the true incidence curve to calculate the lower bounds of the integrated hazards that enter as an adjustment factor in the logistic model. How sensitive is the adjustment procedure to misidentification of the distribution used to calculate the mean ages of onset?

To test the sensitivity of the adjustment procedure, we assume five log normal (LN) distribution functions all displayed in panel (a) of Figure 7 and use these (instead of the true log logistic function) to calculate the integrated hazards. The LN distribution functions cover a wide range of age patterns, with the probability of being diabetic by age 100 (in the absence of mortality) ranging from a low of around 0.05 to a high of close to 1.00. The conditional mean ages of onset of diabetes for ages 30 to 100, the values for the resulting integrated hazard, and the frequency distribution of the 25 estimated coefficients from the adjusted model are shown in panels (b), (c), and (d), respectively. Note that despite obvious differences in the conditional mean ages of onset and the associated integrated hazards, estimates from the adjusted models are centered around values that are close to the true value of the coefficient (1.0). The plot does indeed reveal that the choice of distribution for the calculation of the integrated hazards matters, particularly when the one chosen is extreme relative to the underlying one (compare the distribution functions in the first panel). But even an extreme choice does not generate mean errors exceeding 8%. This pales when compared with the biases associated with unadjusted estimates (as large as 40%). Figure 8 is unequivocal on this point: this figure compares the biases when no adjustment is used, when we use each of the five log normal distributions defined above, and when we use the mean estimate from these distributions. The differences are large and, in the absence of any prior knowledge, correcting for risk heterogeneity is always a better strategy than not correcting at all.

In summary, the adjustment procedure yields correct results when the conditional means that serve to calculate adjustment factors are drawn from a distribution function that is similar to the one that underlies the occurrence of the event of interest. And although its results are sensitive to the specification of this distribution, it still performs much better than a naive approach that ignores mortality differentials.

APPLICATION

We evaluate the adjustment procedure using two data sets on elderly people, one for Mexico, MHAS, and the other for Puerto Rico, PREHCO. Both are panel surveys of elderly populations (50 and above in MHAS and 60 and above in PREHCO). Both consist of two waves separated by two years (MHAS) and four years (PREHCO). The first waves were fielded in 2000 and 2002 in MHAS and PREHCO respectively. Both surveys elicited self-reports on diabetes in the first and second waves, and in both cases there is information on interwave mortality. Within the limitations in population panel data of this kind, MHAS and PREHCO provide us enough information to estimate mortality differentials between diabetics and non-diabetics but not enough to estimate the true incidence of diabetes at adult ages.[22] Below we use observed prevalence data in the first wave (the current status information on diabetes) to estimate the effect of a covariate, education, with and without the adjustment procedure. In addition we retrieve an estimate of the mortality differential between diabetics and non-diabetics: in the absence of information on the true incidence of diabetes, a comparison of the latter with the observed mortality differential is the only benchmark we have to judge the performance of the adjustment procedure.

MHAS

The first column of Table 1 displays estimated effects of a dummy variable for education (D) distinguishing low education (D=1) and high education (D=0) on the observed prevalence of diabetes in the first wave.[23] The second column displays estimates of a logistic model including two controls for the integrated hazard, one for each education group. As one would expect if there is risk heterogeneity, the unadjusted effect of age is negative and the one for education is close to zero. After the adjustment, the effect of age changes sign and the one for education increases (from .032 to .130), has the expected sign (positive) but is only marginally significant (at $p<.05$). Thus, even though the variable for education does not attain statistical significance, the estimates change in the direction we would expect. We do not know, of course, what the truth is: it may well be that there are no differentials by education in the incidence of diabetes and, contrary to our a priori expectations, the adjusted estimates simply reflect this.

We use two adjustment factors: an integrated hazard for the group with D=1 and another for the group with D=0. Using the observed mortality from the interwave period we fitted Gompertz models for each education group separately and irrespective of diabetes status. We then calculated the integrated hazard using the parameters of the Gompertz model for each education group. Thus the standard mortality pattern used to calculate

[22] Although these panel data can be used to obtain (noisy) estimates of diabetes incidence for any subgroup, we cannot do so before ages 50 (Mexico) or 60 (Puerto Rico). Since diabetes in these countries has a relatively early onset, the observed incidence would be too incomplete to retrieve reliable effects of covariates. Additional information for PREHCO can be obtained from http://prehco.rcm.upr.edu and for MHAS from http://www.mhas.pop.upenn.edu/english/home.htm.

[23] Low education is defined as less than 6 years of schooling and high education as 6 years or more.

integrated hazards is the same within education groups (for diabetics and non-diabetics) but different across education groups.[24] Recall that the regression coefficients associated with the integrated hazard are estimates of the differences in mortality levels between diabetics and non-diabetics, e.g. $(e^{\theta} - e^{\vartheta})$, where the parameters θ and ϑ are measures of the mortality levels. In our case these parameters correspond to the logs of the Gompertz constants. Because of the panel nature of MHAS we can actually calculate these mortality levels directly and compare them with those obtained from the adjusted model. While not a perfect test, this contrast will provide an indication of performance of the adjustment. Among those with low education the *observed* mortality difference between diabetics and non-diabetics is .59 whereas the *estimated* difference is .43 (minus the value of the regression coefficient). Among those with high education the observed difference is .60 while the estimate is .36. Lack of concordance between estimated and observed values is probably due to departures from the assumption of identical mortality patterns and/or from the Gompertz model. While not perfect, the rather close agreement between observed and estimated values of mortality differentials is reassuring and we take it as an indication of the suitability of the adjustment.

In summary, the suggested adjustment leads to changes in estimates that go in the expected direction and while they cannot be interpreted as the true effects, we find confirmatory evidence in the modest differences between estimated and observed mortality differentials between diabetics and non-diabetics.

PREHCO

Table 2 presents analogous results for PREHCO. The first column reveals that, unlike the case for Mexico, the effect of education is statistically significant even before adjustment and that, like in Mexico, the effect of age is negative. After adjustment (second column) the effect of education doubles and becomes strongly significant and, as in Mexico, the effect of age changes sign and becomes positive as expected. The estimates of mortality differentials, however, are more removed from the observed values than was the case in Mexico. Thus, the expected difference in mortality levels between diabetics and non-diabetics is estimated to be .14 among those with low education and .09 among those with high education (Table 2, second column). While their relative magnitudes are as expected (larger among those with low education) the values are small compared with the observed quantities, .88 and .89 respectively. Thus, in this case the adjusted estimates reinforce an inference that could have been made with unadjusted values but is on shaky grounds as the test comparing estimated and observed mortality differentials is not reassuring enough.

SUMMARY AND CONCLUSION

By and large, conventional current status analysis in particular and analyses of prevalence data in general give short shrift to potential errors that arise under risk heterogeneity,

[24] This is a refinement that we can introduce only due to the panel nature of the data.

e.g. when the risk of attrition prior to the time at which individuals' status is assessed, t_s, is not independent of the occurrence/non-occurrence of the event of interest. Through suitable approximations and simulations we show that even under mild conditions defining the regime of 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 risk of attrition. We show that the adjustment performs much better than the naive estimate and under some conditions will always yield unbiased estimates. The adjusted estimates are quite robust to the precise function governing the incidence of the event of interest but even large departures from it will produce estimates that are much closer to the true values than naive, unadjusted estimates.

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 logistic function 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 censoring. The integrated hazard on which the adjustment factor rests cannot be calculated without knowledge of this baseline hazard. This may be unproblematic in cases when the cause of censoring is adult mortality, for what matters 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 applications it may not be so clear what the baseline hazard should look like, let alone what its approximate curvature may be within a particular range of ages or durations. The third route of research is to assess the performance of the adjustment in a broader array of empirical cases and to determine the extent to which resulting estimates lead to correct inferences.

APPENDIX: Sensitivity of SMAM to mortality differentials by marital status

Multistate system for first marriage. To assess the effects of differential mortality we simulate a three state system with one absorbing state and three transition rates. All members of a cohort start in the single state and can then either transit to the married state or to the absorbing state of death. Once an individual is married she either stays there or transits to the absorbing state.

Transition rates for the first marriage process. To model the transition rate from single to married we use the three-parameter Coale-McNeill first marriage function (Coale and McNeill, 1972). The first marriage rates are defined by an accelerated failure time model of the following form: $m(x) = K * G_s((x - a_0)/\sigma)$ where $m(x)$ is the first marriage rate at exact age x, G_s is the standard marriage function (Coale and McNeill, 1972), a_0 is the minimum age at which a significant number of first marriages take place, σ is a scale parameter reflecting the speed of first marriage once it begins, and K is the ultimate proportion of individuals who ever marry so that $1 - K$ is the proportion who remain single. We chose values of a_0 ranging from 10 to 17 in intervals of 1, values σ ranging from 0.5 to 2.5 in intervals of 0.5, and values of K ranging from 0.5 to 1.0 in intervals of 0.1. Altogether we defined 240 first marriage functions and associated risks. To model mortality from age 10 to

60 we use Coale-Demeny's female West mortality model with life expectancies in the range of 40-75 (Coale, Demeny, and Vaughan, 1983). We then define different scenarios according to the size of mortality differential between single and married people. First, we select a life table for married individuals with a life expectancy at age 10 equal to, say $\overset{10}{e}$. Second, we define a set of life tables for single individuals so that their life expectancies at age 10 range from $\overset{10}{e}$ to $\overset{10}{e}$ +15. For each combination of married and single life tables and each of the 240 first marriage functions we calculate SMAM using Hajnal's standard expression and compare its value with the mean age at marriage associated with the first marriage function. The difference between the two is due to mortality differentials between single and married individuals.

REFERENCES

Berk, R.A. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Review of Sociology* 48:386-298.

Coale, A.J., P. Demeny, and B. Vaughan. 1983. *Regional Model Life Tables in Stable Populations.* New York: Academic Press.

Coale, A.J. and D.R. McNeil. 1972. "The Distribution by Age of the Frequency of First Marriage in Female Cohort." *Journal of the American Statistical Association* 67:743-749.

Crimmins, E., Y. Saito, and S. Reynolds. 1997. "Further Evidence on Recent Trends in the Prevalence and Incidence of Disability Among Older Americans from Two Sources: The LSOA and the NHIS." *Journal of Gerontology* 52B(2):S59-S71.

Diamond, I.D. and J. McDonald, 1991. "The Analysis of Current Status Data." Pp. 231-252 in *Demographic Applications of Event History Analysis,* edited by T.J. Trussell, R. Hankinson and J. Tilton. Oxford: Oxford University Press.

Freedman, V.A., E. Crimmins, R. Schoeni, B.C. Spillman, H. Aykan, E. Kramarow, K. Land, J. Lubitz, K. Manton, L.G. Martin, D. Shinberg, and T. Waidmann. 2004. "Resolving Inconsistencies in Trends in Old-Age Disability: Report from a Technical Working Group." *Demography* 41(3):417-441.

Fligstein, N. and W. Wolf. 1978. "Sex Similarities in Occupational Status Attainment: Are the Results Due to the Restrictions on the Sample to Employed Women?" *Social Science Research* 7:197-212.

Greene, W. H. 1981. "Sample Selection Bias as a Specification Error: A Comment." *Econometrica* 49:795-798.

Grummer-Strawn, L.M. 1993. "Regression Analysis of Current Status Data: An Application to Breastfeeding." *Journal of the American Statistical Association* 881:758-765.

Hajnal, J. 1953. "Age at Marriage and Proportions Marrying." *Population Studies* 72(2):111-136.

Hajnal, J. 1965. European Marriage Patterns in Perspective. Pp.101-143 in *Population in History: Essays in Historical Demography*, edited by D.V. Glass and D.E.C. Eversley. London, Edward Arnold.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153-161.

Heckman, J. B. Singer. 1984. "A Method for Minimizing the Impact of Distributional

Assumptions in Econometric Models of Duration Data." *Econometrica* 52:271-320.

Hougaard, P. 2000. *Analysis of Multivariate Survival Data.* New York: Springer.

Hu, Y. and N. Goldman. 1990. "Mortality Differentials by Marital Status: An International Comparison." *Demography* 27(2):233-250.

Jewell, N.P. and M.J. Van der Laan. 2004. "Current Status Data: Review, Recent Developments and Open Problems." *Handbook of Statistics* 23:625-642.

Keiding, N. 1991. "Age-Specific Incidence and Prevalence: A Statistical Perspective." *Journal of the Royal Statistical Society: Series A* 154(3): 371-342.

Keiding, N. 2006. "Event History Analysis and the Cross-Section." *Statistics and Medicine* 25:2343-2364.

Keiding, N., K. Begtrup, T.H. Scheike, and G. Hasibeder. 1996. "Estimation from Current-Status Data in Continuous Time." *Lifetime Data Analysis* 2:119-129.

Keiding, N., C. Holst, A Green. 1989. "Retrospective Estimation of Diabetes Incidence from Information in a Prevalent Population and Historical Mortality." *American Journal of Epidemiology* 130(3): 588-600.

Kisker, E.E. and N. Goldman. 1987. "Perils of Single Life and Benefits of Marriage." *Social Biology* 34(3-4):125-152.

Lin, D.Y., D. Oakes and Z. Ying. 1998. "Additive Hazards Regression with Current Status Data." *Biometrika* 85:289-298.

Little, R. J. A. 1995. "Modeling the Drop-Out Mechanism in Repeated-Measures Studies." *Journal of the American Statistical Association* 90:1112-1121.

Livi-Bacci, M. 1985. "Selectivity of Marriage and Mortality: Notes for Future Research." Pp. 99-108 in *Population and Biology*, edited by N. Keyfitz. Liege, Belgium: Ordina Editions.

Manton, K and X. Gu. 2001. "Changes in the Prevalence of Chronic Disability in the United States Black and Nonblack Population Above Age 65 from 1982 to 1999." *Proceedings of the National Academy of Sciences of the United States of America* 98(11):6354-6359.

Manton, K., X. Gu, and V.L. Lamb. 2006. "Change in Chronic Disability from 1982 to 2004/5 as Measured by Long-Term Changes in Function and Health in the U.S. Elderly Population." *Proceedings of the National Academy of Sciences of the United States of America* 103(48):18374-18379.

Schoeni, R.F., V.A. Freedman, and L.G. Martin. 2008. "Why is Late-Life Disability Declin-

ing?" *The Milbank Quarterly* 86(1):47-89.

Manton, K.G. and E. Stallard. 1981. "Methods for Evaluating the Heterogeneity of Aging Processes in Human Populations Using Vital Statistics Data: Explaining the Black/White Mortality Crossover by a Model of Mortality Selection." *Human Biology* 53:47-67.

Smith, James P. 2007. "Nature and Causes of Trends in Male Diabetes Prevalence, Undiagnosed Diabetes, and the Socioeconomic Status Health Gradient." *Proceedings of the National Academy of Sciences of the United States of America* 104(33):13225-13231.

Sun, J. and J.D. Kalbfleish. 1993. "The Analysis of Current Status Data on Point Processes." *Journal of the American Statistical Association* 88(424): 1449-1454.

Trussell, J. and T. Richards. 1985. "Correcting for Unmeasured Heterogeneity in Hazard Models Using the Heckman-Singer Procedure." Pp. 242-276 in *Sociological Methodology,* edited by N. Tuma. San Francisco: Jossey-Bass.

Trussell, J. and G. Rodríguez. 1990. "Heterogeneity in Demographic Research." Pp. 111 -132 in *Convergent Issues in Genetics and Demography,* edited by J. Adams, D. Lam, A. Hermalin, and P. Smouse. New York:Oxford University Press.

Vaupel, J.W., K.G. Manton, and E. Stallard. 1979. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography* 16:439-454.

Vaupel, J.W. and A. I. Yashin. 1985. "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics." *American Statistician* 39:439-454.

Wooldridge, J. M. 1995. "Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions." *Journal of Econometrics* 68:115-132.

Table 1: Estimates of the Effects of Education on the Probability of Being Diabetic: MHAS.

Estimates[a]	**Model with no Adjustment**	**Model with Adjustment**
Constant	-1.44 (1.19)	-7.71 (3.57)
Log of Age	-.72 (.28)	1.52 (.87)
Dummy of Education[b]	.032 (.084)	.13 (.06)
Integrated Hazards		
Low Education	–	-.43 (.25)
High Education	–	-.36 (.35)
N	7586	7586
Log Likelihood	-3515	-3505

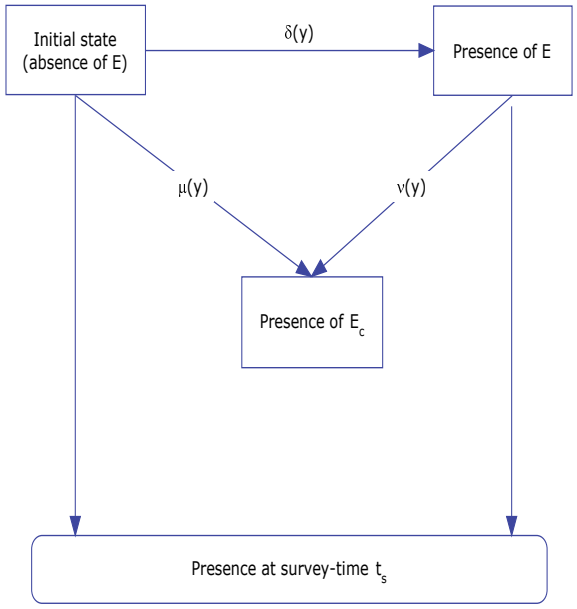
(a) Standard errors in parentheses.
(b) The dummy attains a value of 1 when individuals have less than six years of education, and 0 otherwise.

Table 2: Estimates of the Effects of Education on the Probability of Being Diabetic: PREHCO.

Estimates[a]	**Model with no Adjustment**	**Model with Adjustment**
Constant	2.88 (1.16)	-3.93 (2.61)
Log of Age	-.90 (.27)	.72 (.63)
Dummy of Education[b]	.18 (.07)	.35 (.10)
Integrated Hazards		
Low Education	–	-.14 (.04)
High Education	–	-.09 (.05)
N	5278	5278
Log Likelihood	-3126	-3105

(a) Standard errors in parentheses.
(b) The dummy attains a value of 1 when individuals have less than six years of education, and 0 otherwise.

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



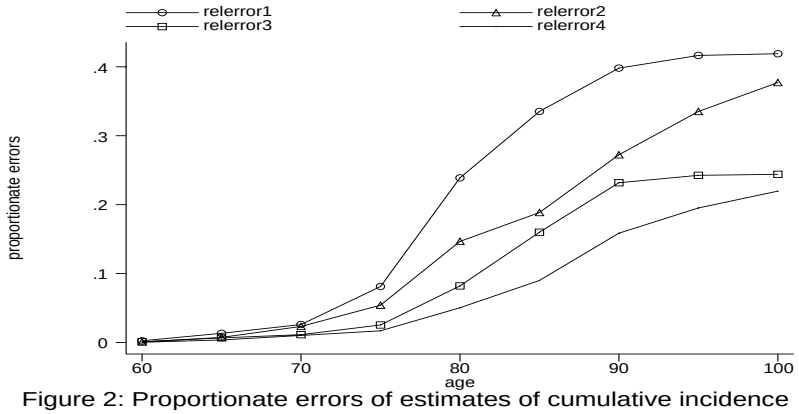


Figure 2: Proportionate errors of estimates of cumulative incidence

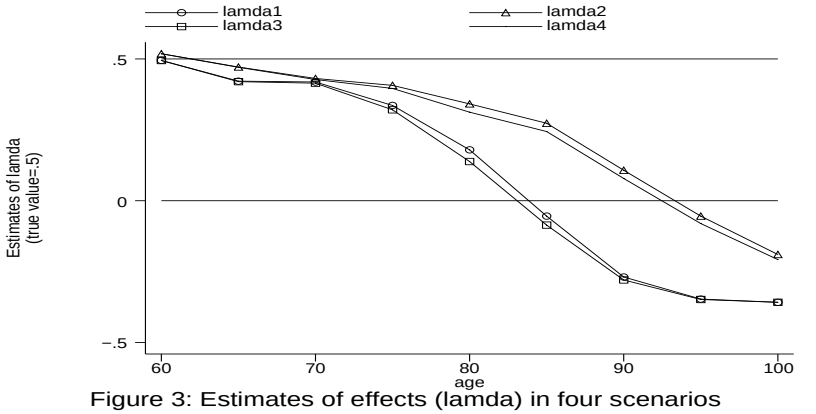


Figure 3: Estimates of effects (lamda) in four scenarios

Figure 4: Simulation of age-specific diabetes prevalence, by education group, in the presence and absence of mortality. The solid lines show the parametric curves used to simulate the data, and the shapes show the results from a single run of the simulation.

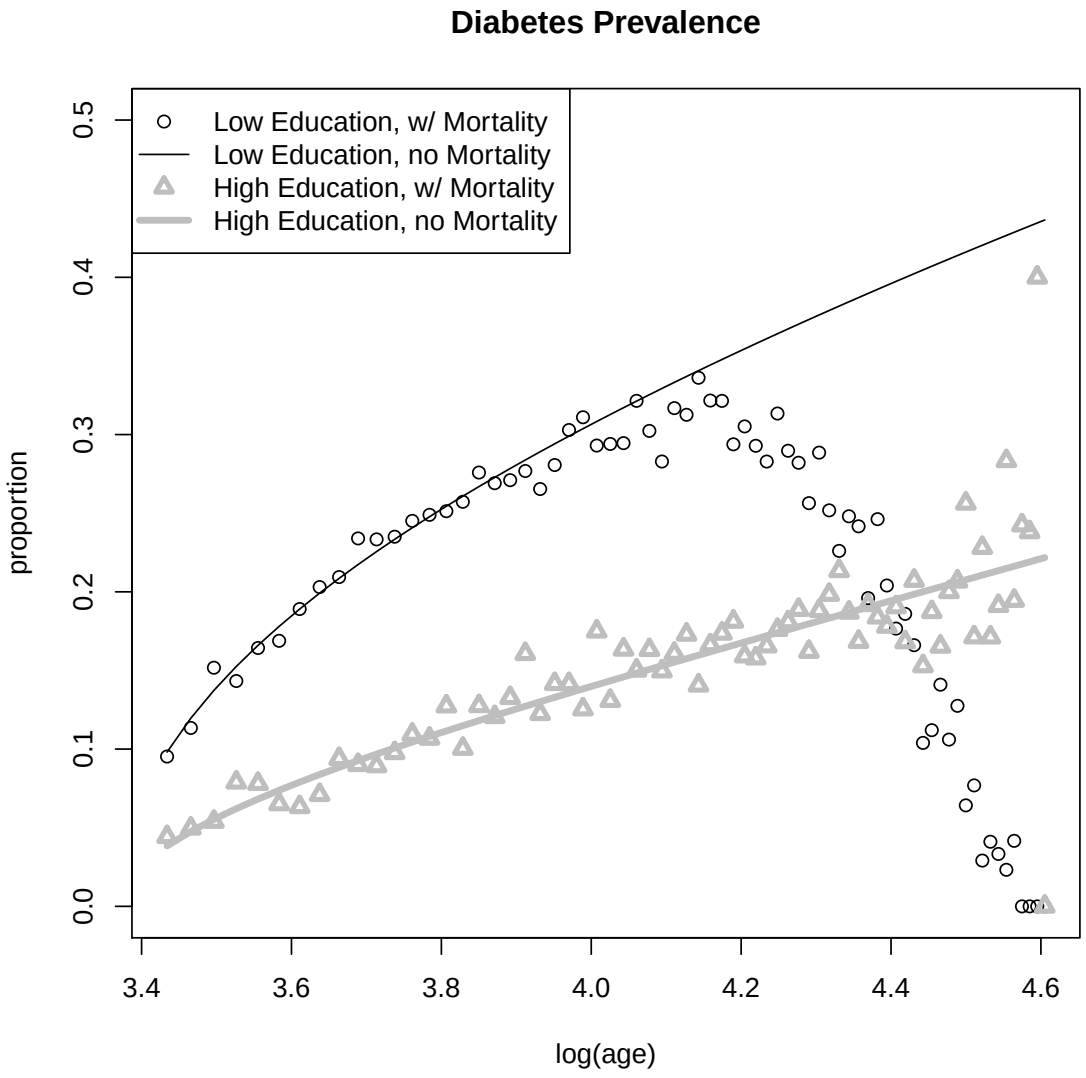


Figure 5: Simulation results: estimated coefficients for low education (dummy variable) in a logit model (of the log odds of being diabetic) for various mortality differentials among the high (top axis) and low (bottom axis) education groups. (Values greater than 1 for the the top and bottom axes indicate higher mortality rates among diabetics for the corresponding education group.) The true value of the coefficient of low education is 1 (horizontal line), and the box plots show the distribution of estimated coefficients over 25 simulations for the corresponding mortality differentials in each education group.

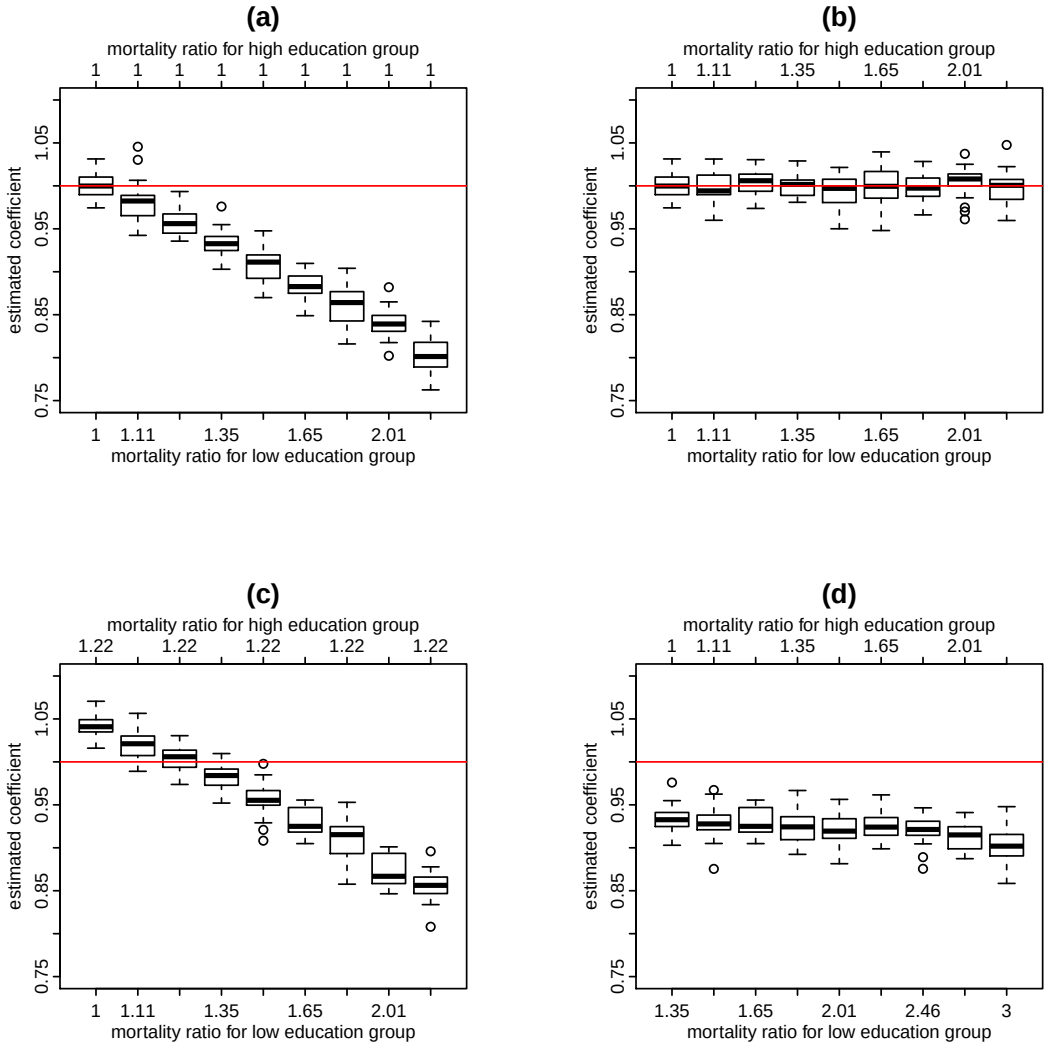


Figure 6: Simulation results: bias (truth - estimate) in the unadjusted and adjusted coefficients for low education (dummy variable) in a logit model of the log odds of being diabetic for various mortality differentials among the high and low education groups. (Values greater than 1 for the axes labeled "mortality ratio" indicate higher mortality rates among diabetics for the corresponding education group.)

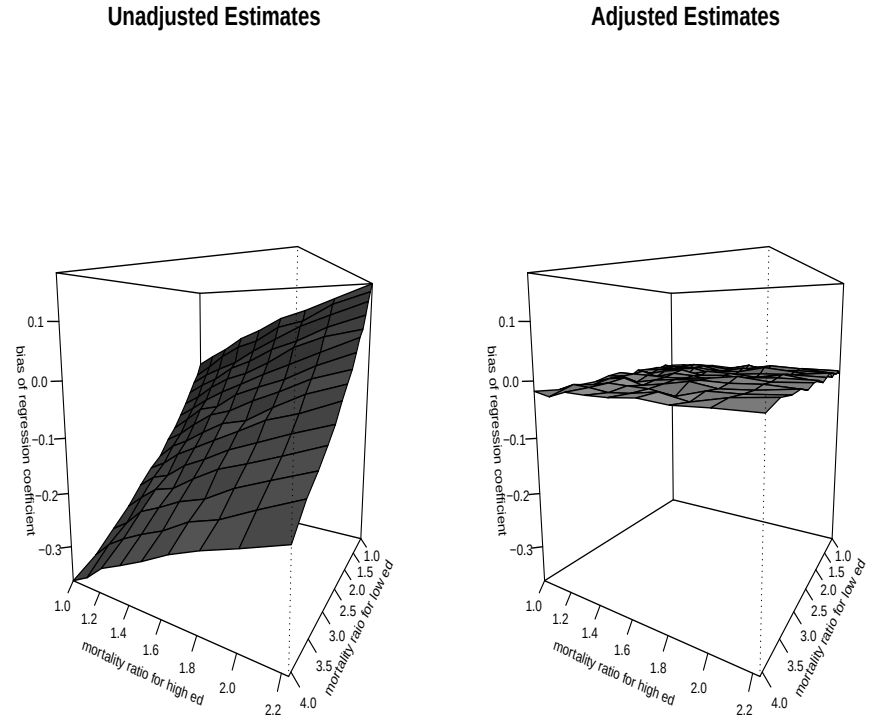


Figure 7: Simulation results: sensitivity of the adjustment procedure to errors in the assumed distribution of diabetes incidence. Characteristics of the true log-logistic distribution are compared to different log normal distributions, LN(mean, sd), in panels a-c. Panel d shows the distribution of estimated coefficients of low education (dummy variable) in a logit model of the log odds of being diabetic for simulated data with only a mortality differential between diabetics and non-diabetics in the low education group.

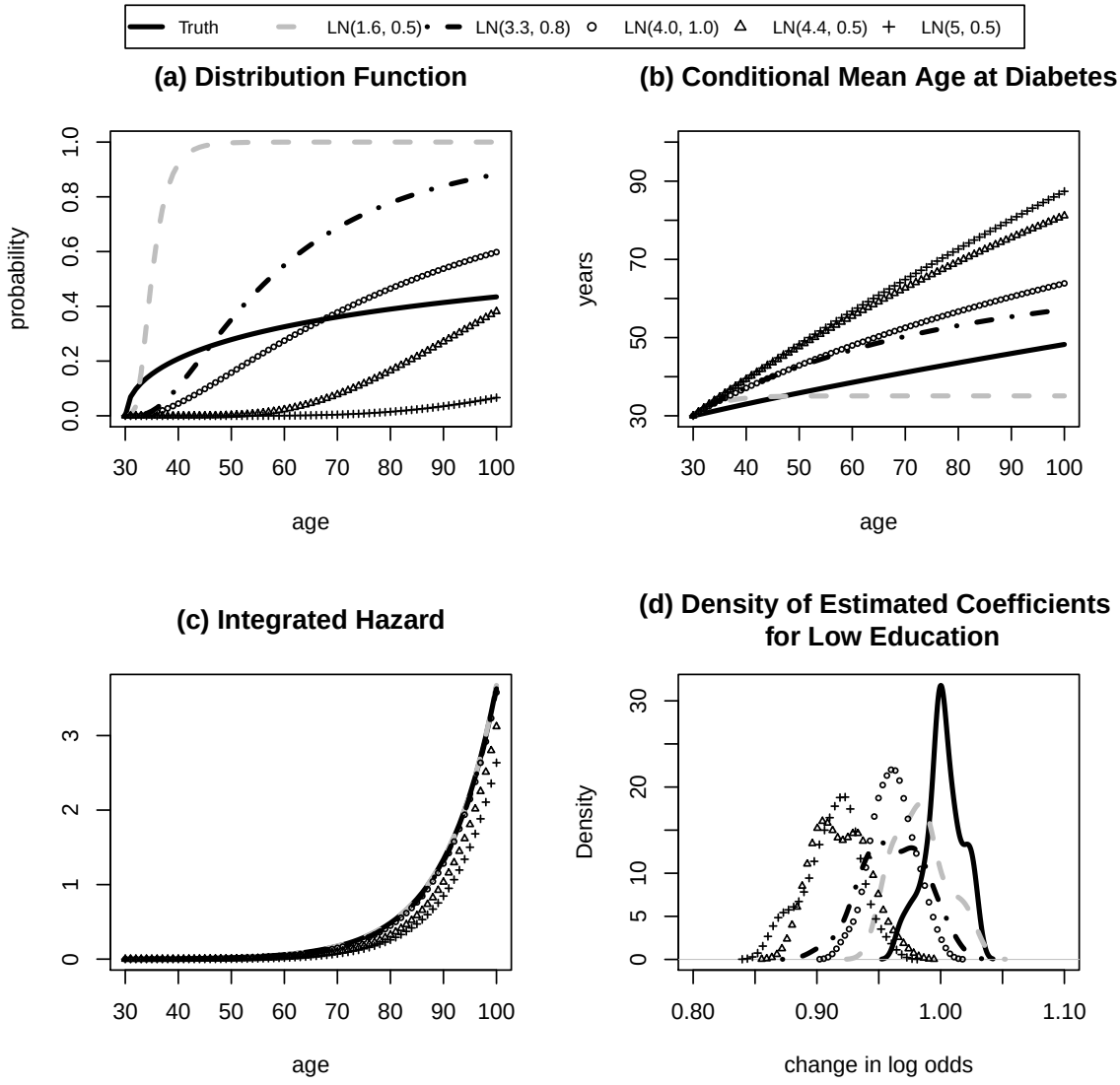
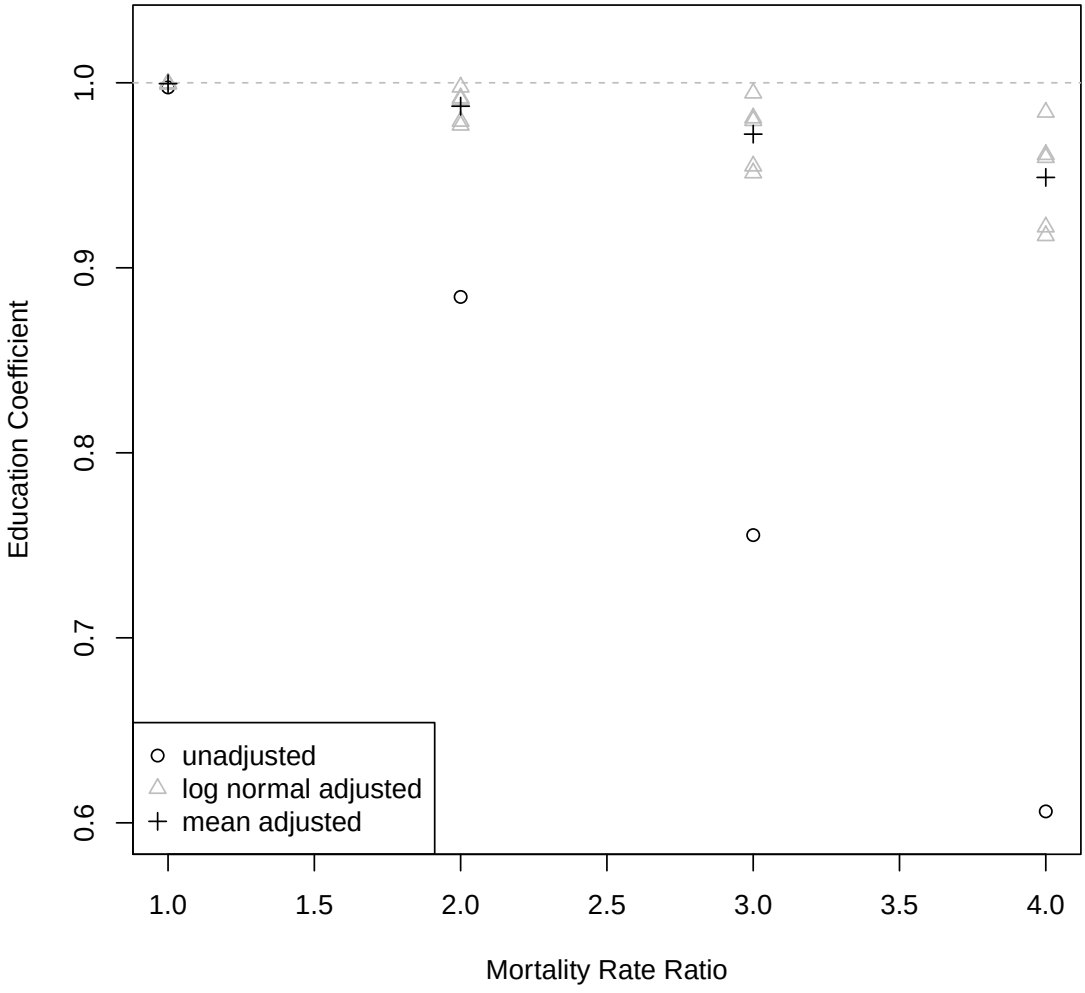


Figure 8: Bias in the Unadjusted, Adjusted, and the Mean of the Adjusted Estimates from the Sensitivity Analysis.



Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: apalloni@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu