ISER Working Paper Series

Estimates of Survival and Mortality from Successive Cross-Sectional Surveys

David W. Smith
London School of Hygiene and Tropical Medicine
Stephanie L. McFall
Institute for Social and Economic Research
University of Essex
Benjamin S. Bradshaw
University of Texas School of Public Health

No. 2010-12
April 2010

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

www.iser.essex.ac.uk

Non-Technical Summary

It is hard to estimate death rates for population groups that are not defined on death certificates. This paper presents a method for estimating death rates for such groups, in this case persons with diabetes, using repeated cross-sectional surveys. The method was originally developed to estimate death rates from repeated national censuses. Survival ratios are the ratio of the estimated number of survivors at a later time to the estimated population size in the initial period. The estimates of population size used in the survival ratios are estimated from two separate and independent surveys. This simplifies the calculation of the variance or degree of dispersion of the survival ratio relative to its average. We illustrate the method using data from the U.S. Behavioral Risk Factor Surveillance System (1996-1998 and 2001-2003) for persons with diabetes. We estimate annual death rates and their degree of precision (standard errors) during the five year period between surveys. Useful estimates of death rates for chronic conditions or other small population groups can be made from sample surveys of the general population when both presence of the condition and age of onset are obtained.

ABSTRACT

Survival ratios and death rates for chronic conditions can be estimated from successive, cross-sectional surveys when the condition and the age of onset are obtained. Survival ratios use the estimated population in the first survey period as the denominator and the estimated number of survivors at a later survey period as the numerator. These ratios have independent numerators and denominators and their variance estimates are a modification of the usual formulas. We illustrate the method by estimating annual death rates and their standard errors among diabetics in the United States.

JEL classifications: I12, C83

KEY WORDS: survival ratios; death rates; chronic disease; diabetes; survey methods, BRFSS

Contact: Stephanie McFall, University of Essex, Wivenhoe Park, Colchester, CO$ 3SQ, UK; tel. +44(0)1206-873897; email smcfall@essex.ac.uk

INTRODUCTION

Mortality rates, primary indicators of health status, are usually reported by age and sex and sometimes other characteristics such as race and ethnicity. These rates are constrained by the items on the records used to compute them: vital records and a census. Mortality rates could also be useful measures of the burden of disease for subpopulations of people who have a chronic disease such as diabetes or a long-term condition such as physical impairment. Such potentially useful categories are rarely recorded on death certificates or obtained in a census. This information is often available from surveys which are used to estimate prevalence rates of diseases or conditions. We describe a method to estimate death rates from successive, independent surveys. This is an adaptation of a method that was developed for survival ratios and death rates calculated from successive national censuses (United Nations, 1967).

To estimate death rates of subpopulations based on health conditions or chronic diseases obtained in surveys, two successive surveys must ascertain whether or not a condition or disease is present at the time of interview. The second survey must also ascertain age at onset or year of onset of the condition to determine if the condition was present when the first survey was done. When this information is available, survival from the time of the first survey to the time of the second can be estimated. The information about age of onset is sometimes available in health surveys since it is used to estimate incidence rates of diseases or conditions (Kirtland, Li, Geiss, et al 2008).

We estimate survival ratios and their variances from two successive, independent surveys and use the survival ratios to estimate death rates and their variances. We illustrate this method by estimating death rates for diabetics in the United States. Diabetes was selected as the case

example because of substantive interest in its impact on population health. Diabetes was a cause of 8% of all deaths in the US in 1999-2001 (Smith and Bradshaw 2008). The prevalence of diabetes, 5% to 10% in recent years, is sufficiently low to challenge the method we use here.

Survey Estimates of Population Sizes

Estimates of the initial size of the subpopulation and its variance can be made using standard survey estimation methods which account for sampling weights and the survey design (Hansen, Hurwitz & Madow, 1953). Estimates of the survivors in the second period require special treatment to account for age at interview and age of onset.

The survivors are the respondents in the subpopulation who would have had their chronic condition status identified by the first survey, that is, who were already diagnosed by that time. Onset of a chronic disease is usually obtained by asking how old a respondent was when they were first diagnosed with the disease or condition, though year of diagnosis or time since diagnosis could be asked instead. The difference between the respondent's reported age at the time of interview and age at diagnosis can be compared with the time since the first survey to determine which respondents had already been diagnosed at that time.

If the two surveys are k years apart, respondents in the second survey whose onset was more than k years before or less than k years before can be easily classified as having had the condition or not at the time of the initial survey. Respondents whose onset is equal to k years before cannot be classified, but their estimated total population size can be divided between the two groups. Since the time of the interview is random throughout the year onset can occur equally before or after a respondent's birthday during the year, on average. Those who were

diagnosed k years since onset could equally have had their onset before or after k years before the second survey so we divide them equally between the two groups.

From the second survey we first estimate the subpopulation sizes of three groups with the given condition by onset time: those with onset before the time of the first survey (Tb), those with onset after the time of the first survey (Ta), and those with onset at the same time as the first survey (Ts). The estimated total of those with onset before the first survey plus one-half the estimated total of those with the condition diagnosed at the same year as the first survey, X = Tb+Ts/2, is the subpopulation surviving from the time of the first survey to the second. (The remainder, the total of one-half of those diagnosed at the same age year of age as the first survey plus the total of those diagnosed more recently, X = Ta+Ts/2, estimates the incident subpopulation after adjustment for deaths occurring after diagnosis.) The variance of the estimated total is V(x) = V(Tb) + V(Ts)/4 + Cov(Tb, Ts). All the components can be estimated with standard survey software that incorporates weights and sample design. These estimates can also be made by categories of age and sex for sex-specific or age-sex specific rates where the sample sizes are sufficiently large.

Survival Ratios and Annual Probabilities of Death

The survival ratio, S = X/Y is the ratio of the number of those surviving, with a previously diagnosed condition, in the second period (X) divided by the number with the condition during the first period (Y). The variance of the ratio estimate can be estimated, treating the numerator and denominator as random variables (Hansen, Hurwitz, and Madow, 1953, vol I, pp. 162-167). They are independent in this ratio since the two surveys are independent. This

simplifies the variance of the estimate somewhat, which is approximately $V(x/y) = (X/Y)^2 (V(X)/X^2 + V(Y)/Y^2)$, since the numerator and denominator are uncorrelated. The standard error is the square root of the estimated variance.

The probability of death during the period between the two surveys is 1 - S. It is useful to estimate annual probabilities since one year is the usual reporting period for probabilities of death while the number of years between surveys, k, can vary. If the annual probabilities of death are constant for the whole period then the annual probability of mortality is $M = 1-S^{1/k}$ where k is the number of years between surveys. The variance of the estimate of M is approximately $V(m) = (S^{1/k -1}/k)^2 V(s)$ using the Taylor series method to obtain this and taking its square root to obtain the standard error. Again, useful estimates can also be made by sex or for age and sex subpopulations where the survey sample sizes are large enough.

Hansen, Hurwitz, and Madow (1953, vol II, pp. 109-111) recommended that the coefficient of variation of the estimated denominator of a ratio statistic be less than 0.05 in order for the usual, approximate, confidence intervals (the estimate plus or minus its standard error times a critical value) to be good approximations to exact, asymmetric intervals derived by Fieller (1940, 1954), which are more accurate for any value of the coefficient of variation.

Estimates for Diabetics in the United States

We estimated survival ratios and annual probabilities of death for a five year period using three years of pooled survey data. The initial period was 1996-1998 and the final period was 2001-2003. Effectively, we treated each pooled estimate as an estimate for the middle year, giving a time interval of five years for deaths to occur, on average.

We used the public use data files of the Behavioral Risk Factor Surveillance System (BRFSS), a large telephone survey sponsored by the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention 2005; Holtzman 2003). The target population is the noninstitutionalized adult population of the United States. Each jurisdiction conducts an independent sample. For 2001-2003, the sample designs were list-assisted with disproportionate stratified sampling (DSS) of telephone numbers, with strata defined by the density of households in the list of numbers. Many states also used geographic strata, primarily to control the sample sizes. For 1996-1998, sample designs were more varied: Mitofsky-Waksberg, DSS, and others.

The response rate routinely reported for the BRFSS is labeled the CASRO (Council of American Survey Research Organizations) response rate. It is the number of respondents divided by the number of in scope units, known units and an estimate of the number in-scope for those of unknown eligibility (Biemer & Larsberg, 2003). For the period 1996-1998 the state median response rates ranged from 59.2% to 63.1%. The minimum state response rate was 32.5% and the maximum was 88.9% (CDC, no date). For 2001-2003, the median response rate for states ranged from 57.1% to 58.3%. The minimum response rate for a state in this period was 33.3% and the maximum was 82.6% (CDC 2002, CDC 2003, CDC 2004).

The core of the BRFSS questionnaire for many years has included the question "Have you ever been told by a doctor that you have diabetes?" The response categories are: yes, no, only while pregnant, don't know or not sure, and refused. We recoded each response as yes or other to compute rates of diagnosed diabetes among all respondents. The BRFSS has an optional module of questions for diabetics which includes "How old were you when you were told you have diabetes?" During 2001-2003 every state but Illinois and Oregon used the optional diabetes

module in at least one year. Our estimates for both periods excluded those two states.

Since age at the time of interview is recorded in years as 18 through 98 with 99 indicating anyone older than 98 we used an initial age range of 18 to 94 and a final age range of 23 to 99, so our estimates apply to diabetics aged 18 to 94. Diabetics who are initially over age 94 and survive at least five years are counted as survivors in our final estimate and slightly increased our estimated survival ratio.

For each period we computed new weights for the pooled samples. For each state we used the original weights and the sample sizes in each of the three years. A respondent's new weight was computed as the original weight times the number of interviews done by the state in the year of the interview divided by the total number of interviews done by the state in all three years. Our reweighting method allows more even weights of respondents in different years, compared with the simplest method of reweighting each year equally, but stops short of complete reweighting by age, sex, region, and other post-stratification categories. For the second period we used only the one, two, or three years of data for each state that included the optional diabetes module. For both estimates we treated states as strata but did not use strata within states.

We also report indirectly standardized mortality ratios for diabetics using the US death rates for 2000 to compute expected deaths by age and sex for five years. The age intervals were 15 to 94 years by 10 years with the first interval providing the estimate for survey respondents aged 18 to 24 years. The death rates for these intervals were weighted by the estimated diabetic population sizes in 1996-1998 to obtain the expected deaths, which were added to get the total expected deaths. The ratio of the survey estimate of the total deaths to the expected number based on US rates is the indirectly standardized ratio for the diabetic population in 1996-1998. The estimated variance and standard error of this ratio used the denominator as a fixed value

though this is subject to sampling variation of the initial sample.

Estimates and derived statistics are shown in Table 1. Of diabetics age 18-94 in the US during 1996-8, 81.4% survived five years, with a standard error (SE) of 1.3%. The corresponding annual death rate was 41.1 per thousand (SE=3.2). This was 2.06 (SE=0.16) times the rate expected for US adults with a similar age-sex composition of the initial sample. Among men the survival ratio was 84.7% (SE = 2.1%) and the annual death rate was 32.8 per thousand (SE = 4.9), or 1.50 (SE=0.22) times the expected rate. Among women the survival ratio was 78.5% (SE = 1.7%) and the annual death rate was 47.3 per thousand (SE = 4.1), or 2.7 (SE=0.23) times the expected rate. All the coefficients of variation of the denominators of the survival ratios are below 0.05.

TABLE 1. Estimated diabetic population in 1996-1998 and estimated survivors in 2001-2003 in the U.S. with survival ratios and annual probabilities of mortality (per 1000). Standard errors (SE) are shown for all estimates and the coefficients of variation (CV) are shown for the initial population estimate.

	Initial Pop'n			Final Survivor		Survival		Annual Prob'y	
Sex	Estimate	SE	CV	Estimate	SE	Ratio	SE	(per 1000)	SE
Male	4,457,101	77,356	0.017	3,773,302	68,761	0.847	0.021	32.8	4.9
Female	5,104,130	73,976	0.014	4,006,599	63,866	0.785	0.017	47.3	4.1
All	9,561,231	106,305	0.011	7,779,901	94,936	0.814	0.013	40.4	3.2

DISCUSSION

It is feasible to estimate death rates for modest sized subpopulations from large surveys conducted several years apart. Estimates of death rates based on surveys appear to be acceptable for planning, as census-based estimates of population mortality have proven their utility (United Nations, 1967). Estimates can be made for characteristics that can be obtained in a survey but are not obtained in a census or from death certificates, such as self-reported chronic diseases or conditions. Systematic collection of the age of onset in surveys would allow estimates of death rates as well as incidence rates for chronic conditions or diseases.

These estimates are subject to the kinds of errors that occur in surveys, including sampling and nonsampling errors. Since the BRFSS does not obtain the information about age of onset each year in every state the final sample sizes in our example were quite variable. This increased the variance of the estimates we made and would also increase the variability of estimates for individual states. Systematic collection of the age of onset would reduce this variability and allow publication of regular rates by states.

The usual estimates of death rates are based on two data sources: death certificates and population estimates from a recent census, each with characteristic sources of error. Estimates of death rates for diabetics that link survey responses with death certificates of respondents have also been used (Gu, Cowie, Harris 1998, Saydah, Eberhardt, Loria, et al 2002) and estimates that use both surveys and death certificates but without linkage of specific records (Tierney, Geiss, Engelgau, et al 2001). These estimates are subject to both sampling and nonsampling errors that differ from both the method we have proposed here and the standard methods for death rates. The choice of method should be influenced by better understanding of the errors in the estimates

as well as the costs and feasibility of alternative methods.

A next step will be to apply this approach to smaller geographic units, states, to examine the impact of smaller sample sizes on the plausibility and precision of estimated mortality as well as provide useful local area estimates. Another step is to apply this method to other chronic conditions where both status and age of onset are obtained in a survey. One example is the Canadian Community Health Survey, which has asked the age of onset of every major chronic condition included in the survey.

REFERENCES

Biemer PP & Lyberg. (2003) Introduction to Survey Quality. New York: John Wiley & Sons.

Centers for Disease Control and Prevention. (no date) 1998 BRFSS Summary Quality Control Report. Retrieved on March 29, 2010: ftp://ftp.cdc.gov/pub/Data/Brfss/98quality.pdf.

Centers for Disease Control and Prevention. (2002) 2001 Behavioral Risk Factor Surveillance System Summary Data Quality Report. Retrieved on March 29, 2010: ftp://ftp.cdc.gov/pub/Data/Brfss/2001SummaryDataQualityReport.pdf.

Centers for Disease Control and Prevention. (2003) 2002 Behavioral Risk Factor Surveillance System Summary Data Quality Report. Retrieved on March 29, 2010: ftp://ftp.cdc.gov/pub/Data/Brfss/2002SummaryDataQualityReport.pdf

Centers for Disease Control and Prevention. (2004) 2003 Behavioral Risk Factor Surveillance System Summary Data Quality Report. Retrieved on March 29, 2010: ftp://ftp.cdc.gov/pub/Data/Brfss/2003SummaryDataQualityReport.pdf.

Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Operational and User's Guide, Version 3.0, March 4, 2005.

Gu K, Cowie CC, Harris MI. (1998) Mortality in adults with and without diabetes in a national cohort of the U.S. population, 1971-1993. Diabetes Care, 21(7): 1138-1145.

Fieller, EC (1940). The biological standardisation of insulin. Journal of the Royal Statistical Society (Supplement). 1:1–54.

Fieller, EC (1954). Some problems in interval estimation. Journal of the Royal Statistical Society B, 16: 175–185.

Hansen MH, Hurwitz WN, Madow WG. (1953) Sample Survey Methods and Theory, vol. I and

II, New York: John Wiley.

Holtzman D. (2003) Analysis and interpretations of data from the US Behavioral Risk Factor Surveillance System. In: McQueen DV, Puska P, Editors. Global Behavioral Risk Factor Surveillance. New York: Kluwer Academic/Plenum Publishers; p. 35-46.

Kirtland KA, Li YF, Geiss LS, Thompson TJ. (2008) State-specific incidence of diabetes among adults -- participating states, 1995-1997 and 2005-2007, Morbidity & Mortality Weekly Report; 57(43): 1169-1173.

Saydah SH, Eberhardt MS, Loria CM, and Brancati FL. (2002) Age and the burden of death attributable to diabetes in the United States. American Journal of Epidemiology, 156(8): 714–719.

Smith DW, Bradshaw BS (2008). Cause-specific mortality rates in chronic disease subpopulations. *The Open Demography Journal*, 1: 11-14.

Stata. (2007) Survey Data, release 10, College Station, TX.

Tierney EF, Geiss LS, Engelgau MM, et al. (2001) Population-based estimates of mortality associated with diabetes: use of a death certificate check box in North Dakota, American Journal of Public Health, 91(1): 84-92.

United Nations. (1967) Methods of Estimating Basic Demographic Measures from Incomplete Data. New York.