JOURNAL OF INSURANCE MEDICINE

J Insur Med 2006;38:105–110

MORTALITY

Computing Exact Excess Death Rates From a Published Mortality Study

Robert M. Shavelle, PhD, MBA; David J. Strauss, PhD, FASA; David R. Paculdo, MPH

We wish to estimate the associated excess death rate (EDR) or mortality ratio (MR) from a published study of persons with a given medical condition. This requires computation of the *expected* mortality in the study population. If age- and sex-specific person years of data are available, this task is straightforward. Most often, however, we have only descriptive statistics — percentage male, average and standard deviation of age — at the beginning of follow up. We show here how this limited information can be used to compute an exact EDR or other quantities of interest.

Address: Life Expectancy Project, 1439 – 17th Ave, San Francisco, CA 94122-3402; ph: 415-731-0240; fax: 415-731-0290; e-mail: Shavelle@LifeExpectancy.com.

Correspondent: Robert M. Shavelle, PhD, MBA.

Key words: Excess death rate, mortality ratio, methodology, age distribution, normal distribution, comparative mortality.

Received: November 18, 2005

Accepted: March 15, 2006

INTRODUCTION

We begin with a very familiar problem. Suppose a published study reports on the follow up of a group of males with disease X, whose average age is 55, and the 10-year survival probability is 70%. How do we estimate the excess death rate (EDR) and/or the mortality ratio (MR), compared to some standard population (perhaps the general population). (See Comment 1 at end of article)

At least two issues have to be dealt with. Firstly, even in the simplest case where all the subjects are exactly age 55, say, at the start of the study, one needs to compute the expected mortality in the general population over the 10 years. Secondly, one must take account of the fact that age 55 is only the average in the study population; the variability around that age proves to have a substantial effect on the estimated EDRs.

The first issue, the expected 10-year survival in the general population, is straightforward in that standard life tables provide the relevant information; but it would nevertheless be helpful to have the computations automated. The second issue, age heterogeneity in the study population, is a little more complicated. In the present article, we show how it can be handled conveniently. The main purpose of the article, however, is to provide readers with a quick and easy way to deal with both problems. As will be explained, we have constructed a publicly available Excel workbook such that the user inputs the relevant

information about the study population, and the required EDRs are immediately provided.

DEALING WITH THE AGE VARIABILITY IN THE STUDY POPULATION

For simplicity, we first consider the case where the follow up is very short, 1 year, and suppose that the death rate is 0.03 (ie, 30 per 1000 per year). If all the study participants were *exactly* 55 years old, there is no problem; one simply refers to standard mortality tables for males of age 55 in the general population. The 2001 US rate[1] is 0.0077, and so the EDR is 0.0300 − 0.0077 = 0.0223, ie, 22 per 1000. The MR is 0.0300/0.0077 = 3.9 = 390%.

However, this does not take into account the variability of ages at the beginning of follow up. To take an extreme case, it may be that 50% of the subjects in the study are 30 years old, and 50% are 80 years old. In that case, the general population mortality is roughly (0.0014 + 0.0725)/2 = 0.0370, leading to very different results. The EDR is then 0.0300 – 0.0370, which is actually negative. This shows that the estimates of the EDR and MR for disease X could be seriously in error if one assumed that all the study subjects were exactly of age 55.

If we know the age of each study participant, then over the short term the expected mortality rate in the general population is the average of the age-specific rates (See comment 2). In practice, however, the complete distribution of ages often cannot be deduced from the published study. How then to proceed?

SINGER'S APPROXIMATION

As far as we can tell, Singer[2,3] was the first to propose a simple approximate solution. He suggested that one add a certain number of years — 3 years was suggested in the cited article as appropriate in many cases — to the average age, 55, and then proceed as if the entire study population was of exact age 55 + 3 = 58. We thus refer to this adjustment as a *Singer correction.* In our simple example, the general population mortality rate at age 58 is 0.0109, so we would estimate the EDR as 0.0191 and the MR as 275%.

Of course, the appropriate Singer correction will not always be 3 years. It depends most critically on the *variation* in the ages in the study population. If there is no variation, each subject being exactly 55 years old, the correction is 0. At the other extreme, where 50% of the subjects are 30 years old and 50% are 80, each subject's age deviates by 25 years from the average (55). In this case, the average mortality (0.0370) is the same as that of males of exact age 73, and the Singer correction is thus (73 – 55) = 18 years.

For simplicity, we have so far only considered the case where the follow-up period of interest is very short. In some cases, the published study may instead specify the percentage of study subjects who survived 10 years. We then need to compare this with the expected average mortality in the general population, also based on a 10-year follow up. Again we may wish to compute a Singer correction, x, such that the 10-year survival probability at exact age 55 + x is the same as in a cohort with the same age distribution as the study population (with *average* age 55).

The appropriate correction will depend on (a) the average age (55 in the above example) and (b) the duration of follow up. The answer could also depend on other factors, such as the mix of males and females in the study and on the assumed *shape* (rather than the dispersion) of the age distribution. Based on our investigations, however, our impression is that these factors are of minor importance.

CALCULATING SINGER CORRECTIONS

We computed appropriate Singer corrections over a wide range of conditions, using a specially constructed Microsoft Excel

workbook. The workbook is publicly available, and we encourage readers to download and use it (see comment 3). The average study population age and standard deviation are specified. By assuming that the actual ages follow a Gaussian (Normal) distribution with these parameters, we computed the proportions of subjects of each possible initial age. Various follow-up periods were considered: 1, 5, 10 and 20 years.

As an example, consider a cohort of males in the general population followed for 10 years. At the beginning of follow up, the average age is 55. The standard deviation of the ages is 10 years, indicating considerable variation. For every possible starting age, we computed the probability of being alive 10 years later (a column of numbers), and computed the weighted average of these probabilities (leading to a single number). The weights (another column of numbers) were the proportions of the cohort at each starting age, according to the Gaussian distribution. We found that the overall 10-year survival percentage was 84.5%. This proved to be the same as the 10-year survival probability for a cohort of persons all of initial age 58. The Singer correction, which is the difference between this age and the average in the cohort (55), is thus 3 years.

The Table shows the appropriate Singer corrections for initial average age 55, Gaussian distribution of ages, and various standard deviations and follow-up periods. Readers interested in exploring these issues are encouraged to download the underlying workbook and adapt it to their interests.

Many medical studies have an average age of roughly 55, and standard deviation of roughly 10. Singer's rule of thumb, 3 years, while not universally applicable, is nonetheless a useful default.

After preparing the previous material, we found that Winsemius[4] had employed a nearly identical approach to the above and generated results similar to those of the Table.

Table. Singer Correction Values for Males of Average Age 55

Standard Deviation	Follow-Up Period (Years)			
	1	5	10	20
0	0.0	0.0	0.0	0.0
3	0.6	0.4	0.3	0.2
5	1.2	1.0	0.9	0.6
10	4.5	4.1	3.5	1.9
20	18.3	12.4	9.1	3.6

In commenting on Winsemius, Singer[5] agreed with the approach and results for a short follow-up period but expressed concern over results for longer-term follow up, such as 10 years. The appropriate corrections for a short follow up and a long one can be appreciably different. Use of a single correction will thus lead to biased estimates of the EDRs (see comment 4). What then shall we do?

A SIMPLE AND EXACT SOLUTION

Recall that our task is to find the EDR or MR associated with a given medical condition, after controlling for the effect of age, over a study period. Our primary emphasis here is on the EDR, though see the following notes.

Suppose, for example, that the study reported 70% survival over a 10-year period. Then we seek the EDR that, when added to every age-specific mortality rate in the general population, yields this same 70% survival. By ''yield'' we mean that one first computes the survival curve for each possible age, then finds the weighted average of these — with weights equal to the proportion of study persons of each age.

The EDR over the k-year period turns out to be $d = -(1/k)\ln[S(k)/\Sigma p_i S_i(k)]$, where k is an integer, ''ln'' is the natural logarithm, p_i is the proportion of study subjects of exact age i, $S_i(k)$ is the (expected) survival function for the baseline cohort of exact age i, and the summation is over i = 0, 1, 2, …, 100. Further

details are given in comment 5. As can be seen, d is merely the difference between the observed mortality rate over the k-year period, $(-1/k)\ln[S(k)]$, and the expected rate, $(-1/k)\ln[\Sigma p_i S_i(k)]$.

Using this formula, we have constructed an Excel workbook to compute the exact EDR. On the same publicly available workbook that was noted earlier (see comment 3), there is a worksheet labeled ''Compute Constant EDR.'' The worksheet has input areas for the average and standard deviation of ages of males and females, the percentage of males, and the baseline (or reference) mortality rates for males and females. One also enters the observed survival percentages for any or all of the 1-, 5-, 10- and 20-year follow-up periods. The exact EDRs are then computed for these time periods (see comment 6).

As an example, suppose the study population was 90% male (10% female) with average age 60 for males and 65 for females, with respective standard deviations of 5 and 10. Suppose also that the appropriate baselines mortality rates are the 2001 US male and female general population.[1] Lastly, suppose that the study reported a 10-year survival probability of 75%. This is an annual mortality rate of $(-1/10)\ln(0.75) = 0.0288$. According to the workbook, the 10-year expected survival probability for the composite baseline group is 80.9%, for an annual mortality rate of $(-1/10)\ln(0.809) = 0.0212$. The EDR over the 10-year period is thus $0.0288 - 0.0212 = 0.0076$, which is what the workbook shows.

PRACTICAL CONSIDERATIONS

1. As in all work of this type, one must first specify the appropriate baseline mortality rates (or ''expected mortality''). For illustration we have used the general population here. As has been noted (eg, see Singer[2]), the choice of baseline rates often has a larger effect on the computed EDR or MR than any other assumption.
2. We indicated that the user is to select the follow-up period of interest, k years. For example, if the study population is followed for 15 years, one might be interested not only in the EDR commensurate with the observed 15-year survival percentage, but also those based on, say, the 3 periods 0–5, 5–10 and 10–15 years. It is conceivable that the EDRs for the 3 periods show a pattern, such as increasing with duration, rather than appearing constant. Accordingly, the workbook also shows *conditional* survival probabilities, so that the user can easily compare EDRs over these periods, and then decide if they are sufficiently close to merit a constant value.
3. More generally, one need not restrict to a constant EDR for both sexes and all ages. If separate study survival figures are given by sex or starting age, one could estimate these strata-specific EDRs. Alternatively, under the assumption that there is no duration effect, one could specify an age-dependent model for the EDRs. For example, one could model the EDRs to increase linearly with age or in inverse proportion to the remaining life expectancy.[6] Or, one could specify that the female EDRs be exactly half the male values. In such cases, a closed-form solution for the EDR may not exist. However, one can easily compute the EDR by trial and error, either computerized or manually (see comment 7). In the workbook, we have included a worksheet called ''Compute Variable EDR'' for this task. For the reader who would like to explore the methods or results indicated here, we also provide a worksheet labeled ''Check EDR,'' which is the reverse of the above (see comment 8).
4. Thus far, we have only considered estimation of the EDR. It would be equally possible to consider the MR (see comment 9), though the latter summary figure is more dependent on the chosen baseline rates and is thus less generally applicable.

5. Some published studies stratify subjects by age group and provide the proportions in each. If so, this age distribution should be used. In the workbook it is easy to do. One simply specifies the distribution in the column of probabilities instead of using the default Gaussian distribution.

SUMMARY AND CONCLUSION

One often needs to estimate the EDR (or MR, etc) from a published follow-up study of persons with a given medical condition. The 4 necessary steps are:

1. Use the published survival probabilities to compute the mortality rates in the study population.
2. Decide upon the reference population of interest (eg, general population or insured population).
3. Compute the expected mortality in the reference population.
4. Find the EDR as the difference between the rates in steps 1 and 3.

Step 3 is complicated by the variability in the starting ages in the study population; the greater the variability in starting age, the higher the overall mortality rates. This is because mortality in the older subjects is very high and more than outweighs the low mortality of the younger subjects. When computing rates in the reference population, one must take into account the age variability.

Singer[2,3] suggested a useful approximation. In some cases, one can take account of the variability simply by adding 3 years to the mean starting age in the published study. However, it is both easier and more accurate to use the electronic workbook we have provided, which automates steps 1, 3 and 4. The example given here illustrates the process.

We welcome feedback from users.

The authors thank Dr. Richard Singer for correspondence on this topic, and Pierre Vachon for helpful discussion.

REFERENCES

1. Arias E. United States Life Tables, 2001. *National Vital Statistics Reports*. Vol 52, No 14. Bethesda, Md: National Center for Health Statistics; 2004.
2. Singer RB, Kita MW. Guidelines for evaluation of follow-up articles and preparation of mortality abstracts. *J Insur Med*. 1991;23:21–29.
3. Singer RB. The application to life table methodology to risk appraisal. In: RDC Brackenridge, WJ Elder, eds. *Medical Selection of Life Risks*. 4th ed. New York, NY: Stockton Press; 1998:52–53.
4. Winsemius DK. Improved calculations of group mean expected mortality rates, part 1: The case of normally distributed ages. *J Insur Med*. 2000; 32:5–10.
5. Singer RB. Commentary on improved calculations of group mean expected mortality rates, part 1: The case of normally distributed ages. *J Insur Med*. 2000;32:93–95.
6. Strauss DJ, Shavelle RM. Life expectancy of persons with chronic disabilities. *J Insur Med*. 1998;30:96–108.

COMMENTS

1. We are assuming for simplicity that the study population, apart from having disease X, is otherwise comparable to the general population. If not, then one would choose a different reference group, such as an insured population.
2. This approximation only gives reasonable results in practice if the follow-up period is very short (less than a few years) or the strata-specific rates are very similar. In general it is first necessary to compute the expected survival curve for each age stratum and take a weighted average, the weights being the proportions of subjects in each stratum. One can then compute annual mortality rates from this aggregate survival curve.
3. The workbook is available at http://www.LifeExpectancy.org/articles/edr.shtml.
4. For example, consider a population of males of average age 55 with standard deviation 10. Suppose that the true EDR is 0.010. It can be shown that use of a cohort of age 55 + 4 = 59 will lead to the

following estimates of the EDRs: 1-year, 0.011; 5-year, 0.010; 10-year, 0.009; 20-year, 0.005.

5. For simplicity, consider a unisex population. Let p_i be the proportion of persons in the study population of exact age i, and the baseline (general population) mortality rate at age i be denoted by m_i, i = 0, 1, 2, …, 100. Note that the probability that a person of age i will survive exactly k additional years is $S_i(k) = \exp[-\Sigma m_j]$, where k is an integer and the summation is of k terms, from j = i to (i+k−1). Let S(k) denote the composite survival experience for the entire group. Clearly, $S(k) = \Sigma p_i S_i(k)$, where the summation is over i = 0, 1, 2, …, 100. Now consider when all the mortality rates are increased by an EDR, d. $S(k) = \Sigma p_i \exp[-\Sigma(m_j+d)] = \Sigma p_i \exp[-\Sigma m_j] \exp[-\Sigma d] = \Sigma p_i S_i(k) \exp[-kd] = \exp[-kd] \Sigma p_i S_i(k)$. Solving for the EDR gives $d = (-1/k) \ln[S(k)/\Sigma p_i S_i(k)]$. The solution for the more general case, of a population of males and females, is $d = (-1/k) \ln[S(k) / \{\pi \Sigma p_{mi} S_{mi}(k) + (1-\pi) \Sigma p_{fi} S_{fi}(k)\}]$, where π is the fraction of males in the study and the subscripts m and f denote distinct values of p and S for the two sexes.
6. The age, sex and survival percentage inputs are shaded yellow, at the top. The mortality rates are in rows 5 through 105 of columns K and U. The outputs, computed EDRs over the time period and intervals, are shaded purple.
7. Excel has a function called ''Goal Seek,'' accessible on the Tools menu, which essentially performs trial and error. Alternatively, the user can vary the value in the cell for the EDR until the calculated values match the desired observed figures.
8. The worksheet generates two sets of survival distributions, one with a user-specified excess death rate (labeled ''EDR = d'') and one without (''EDR = 0''). It then computes the composite survival percentage (that is, weighted by the starting age distribution) for the two groups using the same baseline mortality rates. One group has baseline mortality rates increased by the specified EDR; the other does not. Based on these values, the worksheet computes the EDRs over 1-, 5-, 10- and 20-year periods. As can be seen, these are each identical to the specified EDR.
9. In estimating the mortality ratio, r, the equation in comment 5 above reduces only to $S(k) = \Sigma p_i [S_i(k)]^r$. Thus, we do not find a simple equation for r. The value of r, however, can be found using computerized trial and error; a modified version of the provided ''Compute Variable EDR'' worksheet could be used for this purpose.