eAppendix: indirect estimation of chronic disease excess mortality and its associated uncertainty using cross sectional data

Pieter HM van Baal, Rudolf T Hoogenveen, Peter M Engelfriet & Hendriek C Boshuizen

Derivation of the formula expressing excess mortality rates as a function of incidence and prevalence

The goal was to derive an equation expressing excess mortality rates, defined as the difference in mortality between persons with and without the disease, as a function of incidence and prevalence. The first step is to describe the survival over time of a cohort consisting of both persons with and without a particular chronic disease:

$$\frac{dS(t)}{dt} = -S(t,d) * m(t,d) - S(t,\bar{d}) * m(t,\bar{d}) \qquad (1)$$

$S(t)$	*Probability of being alive at time t*
$S(t,d)$	*Probability of being alive at time t and having disease d*
$S(t,\bar{d})$	*Probability of being alive at time t without having disease d*
$m(t,d)$	*Mortality probability for those with disease d at time t*
$m(t,\bar{d})$	*Mortality probability for those without disease d at time t*

Equation (1) simply states that the change in the probability of survival over time is the negative of the sum of the probabilities of dying for those with and those without the disease. If we now focus on the part of the cohort with the disease we can state the following:

$$\frac{dS(t,d)}{dt} = i(t) * S(t) - S(t,d) * m(t,d) \qquad (2)$$

$i(t)$	*incidence rate for disease d at time t*

It should be noted that the incidence rate in accordance with the data model equation (2) refers to the total population, not to the disease-free population. It may seem illogical that incidence refers to the whole population (i.e. including those with the disease). However, this is the manner in which most cross sectional data are presented. The probability of having the disease (conditional upon survival) can now be written as:

$$p(t) = \frac{S(t,d)}{S(t)} \tag{3}$$

$p(t)$ *Probability of having disease d at time t conditional on survival*

Now, we will show that by using the derivative of p(t) (that is, the change in the probability of having disease d conditional on survival), we can express the difference in mortality between those with and those without a particular disease as a function of p(t) and incidence only:

$$\begin{aligned} \frac{dp(t)}{dt} &= \frac{\frac{dS(t,d)}{dt} * S(t) - S(t,d) * \frac{dS(t)}{dt}}{S(t)^2} \\ &= \frac{\frac{dS(t,d)}{dt} - p(t) * \frac{dS(t)}{dt}}{S(t)} \end{aligned} \tag{4}$$

From (1) and (2) we have:

$$\begin{aligned} &\frac{\frac{dS(t,d)}{dt}}{S(t)} = i(t) - m(t,d) * p(t) \\ &\frac{p(t) * \frac{dS(t)}{dt}}{S(t)} = -p(t) * [p(t) * m(t,d) + \{1 - p(t)\} * m(t,\bar{d})] \end{aligned} \tag{5}$$

Thus, we can write (4) as:

$$\begin{aligned} \frac{dp(t)}{dt} &= i(t) - m(t,d) * p(t) + p(t)^2 * m(t,d) \\ &\quad + p(t) * m(t,\bar{d}) - p(t)^2 * m(t,\bar{d}) \\ &= i(t) + [p(t)^2 - p(t)] * [m(t,d) - m(t,\bar{d})] \\ &= i(t) - p(t) * [1 - p(t)] * [m(t,d) - m(t,\bar{d})] \end{aligned} \tag{6}$$

So we obtain:

$$[m(t,d) - m(t,\bar{d})] = \frac{i(t) - \frac{dp(t)}{dt}}{p(t) * [1 - p(t)]} \tag{7}$$

We will take this difference as our definition of the excess mortality (η) associated with a particular disease:

$$\eta(t) = [m(t,d) - m(t,\bar{d})] \quad (8)$$

Equation (7) shows that excess mortality rates can be expressed as a function of incidence and prevalence if we know how these quantities develop over time in a cohort. However, our goal is to apply this equation with cross-sectional data on incidence and prevalence specified by age. This does not need to be an obstacle to using our equations if the data are derived from a population that is in a steady-state. In a steady state, we may assume that age-specific incidences and prevalences (and mortality rates) remain constant. We can then assume (as is done in the life table method) that the age-specific data that are available from cross-sectional studies actually represent the changes an aging individual experiences over time. Thus, we may consider the age variable as a proxy for the time variable:

$$\begin{aligned} i(t) &= i(a) \\ p(t) &= p(a) \\ \eta(a) = \eta(t) &= \frac{i(a) - \frac{dp(a)}{da}}{p(a) * [1 - p(a)]} \end{aligned} \quad (9)$$

Equations 1–9 can now be interpreted as a description of a cross-sectional population, implying an assumption of steady state for the disease.

Data used in the example

Data for our example study were incidence and prevalence counts from 4 different general practitioners (GP) registries in the Netherlands. By linking together electronic medical records, several general practice registration networks have been established which can be used for epidemiological studies. The data utilized in this study comes from four of these networks, which are considered to be nationally representative. For each of these 4 registrations the following data were available for the year 2007: incidence and prevalence counts for chronic heart failure (CHF), population size as of January 1, as well as the number of personyears of the total population in the GP registration. All data were grouped by sex and 10 age classes (40-45, 45-49,….., 85+).

Regression modeling strategy

Generalized linear mixed models were estimated describing incidence and prevalence as a function of orthogonal polynomials of age. Incidence and prevalence were estimated simultaneously as a multivariate outcome. This can be easily realised in standard software by combining all data and fitting a model using a dummy variable indicating whether the

outcome is incidence or prevalence. We estimated this multivariate outcome assuming a binomial distribution and a logit link function:

$$E[y(a)] = \frac{1}{1+\exp[-\{\alpha_0 + \sum_{=1}^{n} a^j \alpha_j + \beta_0 I + \sum_{i=1}^{m} a^i \beta_i I\}]} \qquad (10)$$

where

y *dependent variable (incidence or prevalence)*

I *Indicator variable that has value 1 if the dependent variable is incidence and 0 if the dependent variable is prevalence*

We estimated cumulative incidence probabilities using population size instead of person years. This assumption could be made, as GP registrations represent fairly stable populations. We transformed the probabilities into rates by assuming that incidence occurred on average halfway during the year. To capture the possible systematic differences between the four GP registrations the registration identifier was entered into the models as a random intercept. To select the 'optimal' regression models in terms of highest order polynomial we used the Bayesian Information Criterion.[1] Since we used our regression model to indirectly estimate the parameters that describe the excess mortality rates we opted for the BIC criterion since it penalized inclusion of additional parameters more than the AIC criterion. Predictions for incidence and prevalence for the 'average' GP registration were entered as values into equation (9) to estimate excess mortality rates. This was accomplished by fixing the 'random' intercept at its average value. Life expectancy with CHF was estimated by decomposing population mortality rates for 2007 from Statistics Netherlands into mortality rates for persons with and without CHF:[2]

$$\begin{aligned} m(a,\bar{d}) &= m(a) - p(a) * \eta(a) \\ m(a,d) &= m(a,\bar{d}) + \eta(a) \end{aligned} \qquad (11)$$

$m(a)$ *Mortality probability in the total population at age a*

Uncertainty assessment for excess mortality and life expectancy was done using Monte Carlo simulation. With the estimated coefficients and their covariance matrix from the optimal regression models, we set up a Monte Carlo simulation to calculate random excess mortality rates and life expectancy with CHF. Sample size was set at 5,000. Modeling incidence and prevalence as one outcome variable takes into account the correlation between incidence and prevalence. Such a correlation is expected, as high incidences will lead to high prevalence rates, and GP-registrations with high incidences can be expected to have high prevalence rates

too. By modeling incidence and prevalence simultaneously, it is possible to let incidence and prevalence share the GP specific random effect in the regression model and to quantify the correlation between parameters that describe incidence and prevalence.

Exploring the effect of possible time trends

Sensitivity analyses were directed towards testing the key assumption of a 'steady state'. We calculated excess mortality rates and life expectancy in scenarios in which we assumed an annual increase or decrease in the prevalence of respectively 1%, 5% and 10% in our heart failure example. Thus, besides the age related increase there is change in the prevalence proportion due to time effects (e.g. if in year 1 prevalence at ages 50 and 51 is 0.2 and 0.21 respectively, in year 2 these will be .22 and .231 assuming a 10% time trend. The increase in prevalence within a year of a cohort aged 50 years in year 1 will then be .231 - .2 = .031). Equation (9) reveals that not accounting for a positive time trend will result in an underestimate of excess mortality and an overestimate of disease duration. Table 1 displays results of sensitivity analyses. It can be seen that results are sensitive to time trends. However, it should be noted that a 10% annual increase in age specific prevalence is large, since it would imply an almost 50% increase of the incidence.

Table 1: Estimates of disease duration (life expectancy with disease) for the average male CHF patient in the Netherlands in 2007 (These were obtained by weighing age specific estimates of disease duration by prevalence numbers.)

Annual time trends in disease prevalence	Average disease duration
- 10%	3.3
-5 %	3.8
-1 %	4.2
0	**4.3 (3.9 – 4.9)**
+ 1%	4.5
+ 5%	5.2
+ 10%	6.5

References

1. Schwarz GE. Estimating the dimension of a model. *Annals of Statistics* . 1978;6:461–464.
2. Statistics Netherlands. STATLINE. www.statline.nl