ESTIMATING INCIDENCE FROM AGE-SPECIFIC PREVALENCE IN GLAUCOMA

M. CRISTINA LESKE,1 FRED EDERER1 AND MARVIN PODGOR1


A simple method is presented to estimate incidence from age-specific prevalence data for diseases that are irreversible and do not affect mortality risk. The application of the model is illustrated with data on primary open angle glaucoma.

Materials and methods

Both incidence and prevalence are important measures of disease frequency. Data on prevalence are generally easier to obtain than on incidence: only a survey is needed to measure prevalence, while it is necessary to follow a cohort to generate incidence data. Thus, in many situations, data may be available on prevalence (existing cases at a point in time) but not on incidence (new cases developing during a time period). In these situations, is it possible to estimate incidence from age-specific prevalence? In general, the answer is negative, because prevalence is a function of both incidence and duration of disease, and there is generally no simple relationship between these three factors. However, diseases that are both irreversible and nonfatal such as glaucoma, may be an exception: for such diseases, it may be possible to estimate incidence from age-specific prevalences. This paper presents a method for making such estimates, using primary open angle glaucoma (OAG) as an example.

Materials and methods

Age-specific prevalence data on glaucoma

The term glaucoma encompasses several clinical entities which are characterized by abnormal intraocular pressure, cupping of the optic disc, and nerve bundle fiber defects of the visual field. The most common type is OAG, one of the main causes of blindness in the United States (1). Although the disease lacks a standard definition, for purposes of this report, an essential diagnostic criterion is the presence of characteristic glaucomatous visual field loss (2).

Among the few published general population prevalence surveys of glaucoma, two have used visual field loss as a diagnostic criterion. One of these studies screened 92 per cent of the 4608 residents of three Welsh villages who were 40–75 years of age (3). Twenty persons met the diagnostic criteria for OAG, which were based on the presence of optic disc cupping and glaucomatous visual field defects. Age-specific prevalence in 5-year age groups, calculated from their data, is presented in table 1.
ESTIMATING INCIDENCE FROM AGE-SPECIFIC PREVALENCE

607

TABLE 1
Age-specific prevalence of open angle glaucoma in Wales, 1966, and Framingham, 1978 (excluding blind spot enlargement)

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Wales (3) Cases/population</th>
<th>Framingham* (5) Cases/population</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-44</td>
<td>0/792 = 0.0000</td>
<td></td>
</tr>
<tr>
<td>45-49</td>
<td>0/760 = 0.0000</td>
<td></td>
</tr>
<tr>
<td>50-54</td>
<td>2/804 = 0.0025</td>
<td></td>
</tr>
<tr>
<td>55-59</td>
<td>6/718 = 0.0084</td>
<td>3/625 = 0.0048</td>
</tr>
<tr>
<td>60-64</td>
<td>3/686 = 0.0044</td>
<td>5/545 = 0.0092</td>
</tr>
<tr>
<td>65-69</td>
<td>5/495 = 0.0101</td>
<td>4/434 = 0.0092</td>
</tr>
<tr>
<td>70-74</td>
<td>4/353 = 0.0113</td>
<td>6/353 = 0.0170</td>
</tr>
<tr>
<td>75-79</td>
<td></td>
<td>5/258 = 0.0194</td>
</tr>
<tr>
<td>80-84</td>
<td></td>
<td>4/137 = 0.0292</td>
</tr>
<tr>
<td>Total</td>
<td>20/4608 = 0.0043</td>
<td>27/2352 = 0.0116</td>
</tr>
</tbody>
</table>

* Available Framingham data for the age group 52-54 years were excluded.

The other survey was conducted as part of the Framingham Eye Study (4), which examined 2477 (or 84 per cent) of the survivors, aged 52-85 years, of the Framingham Heart Study still living in the area. The diagnosis of glaucoma was based on the presence of rigorously defined field defects and one or more of the following: abnormal cup/disc ratios, abnormal intraocular pressure, and history of glaucoma. Visual field defects, in addition to the typical glaucomatous types, also included blind spot enlargement. There were 27 cases of OAG, excluding blind spot enlargement, among persons in the 55-84 year age range, as shown in table 1 (5). Despite differences in diagnostic criteria and case-finding methods, both studies found a similar per cent prevalence in the 55-74 year age range: 0.80 in Wales and 0.92 in Framingham.

Estimating incidence from prevalence

The incidence of glaucoma has been determined for selected groups (6), but not for a general population. In order to estimate glaucoma incidence from general-population age-specific prevalence data, we assume the following: 1) the duration of glaucoma is life-long after diagnosis, since the disease is considered irreversible; 2) the mortality risk is the same in glaucomatous and non-glaucomatous individuals; and 3) glaucoma is a stable disease (7) in a stable population, i.e., disease incidence and population composition (with respect to glaucoma risk factors) remain constant over time.

Under assumptions 1 and 3, which are similar to those of Muench (8), the number of existing cases at the end of an age interval will equal a) the prevalent cases at the beginning of the interval and alive at the end of the interval, plus b) the incident cases occurring during the interval and remaining alive until the end of the interval.

This can be expressed as:

\[ N_{x+1}P_{x+1} = N_xP_x(1 - q_x) + \Pi_x(N_x - N_xP_x)(1 - q_x), \]

where

\[ N_x = \text{population at the beginning of age interval } x; \]
\[ P_x = \text{prevalence proportion at beginning of age interval } x, 0 \leq P_x \leq 1; \]
\[ q_x = \text{probability of dying during age interval } x, 0 \leq q_x \leq 1; \]
\[ \Pi_x = \text{probability of developing the disease during age interval } x, 0 \leq \Pi_x \leq 1; \]
\[ x+1 \text{ designates the age interval immediately following age interval } x. \]

Then

\[ \Pi_x = \frac{N_{x+1}P_{x+1} - N_xP_x(1 - q_x)}{(N_x - N_xP_x)(1 - q_x)}. \]

According to assumption 2, \( N_{x+1} = N_x(1 - q_x) \), thus

\[ \Pi_x = \frac{P_{x+1} - P_x}{1 - P_x}. \] (1)

In a disease of low prevalence such as glaucoma, the denominator in equation 1 is essentially equal to unity, therefore a simple subtraction of consecutive age-specific prevalence proportions closely approximates \( \Pi_x \). As defined above, \( \Pi_x \) is a probability and not a rate and is different from the incidence rate \( I_x \), in
the same way that in a life table, $q_x$ is different from $M_x$, the mortality rate (9). In addition, $\Pi_x$ is a net risk, i.e., the probability of disease in the absence of the competing risk of death, while $I_x$ is a crude rate—the incidence rate in the presence of competing mortality (10). It is possible, however, to estimate $I_x$ from $\Pi_x$. $I_x$ is the number of cases of disease developing in interval $x$ divided by the average population at risk (9). The numerator of $I_x$ is comprised of the new cases that survived and those that did not survive the interval. The number of surviving new cases is $\Pi_x(N_x - N_x P_x) (1 - q_x)$, the product of the probability of disease, $\Pi_x$, and the number of susceptibles surviving the interval, $(N_x - N_x P_x) (1 - q_x)$. However, the number of new cases dying in the interval is only a fraction of $\Pi_x(N_x - N_x P_x) q_x$. The expression $\Pi_x(N_x - N_x P_x) q_x$ is the product of the probability of disease, $\Pi_x$, and the number of susceptibles dying in the interval, $(N_x - N_x P_x) q_x$, as such, it includes those instances of death intervening both before the occurrence of disease (deaths among non-cases) and after the occurrence of disease (deaths among cases). Only the latter will be included in the numerator of $I_x$. These new cases dying in the interval can be expressed as $\Pi_x(N_x - N_x P_x) q_x f_x$, where $f_x$ is the fraction of $\Pi_x(N_x - N_x P_x) q_x$ where disease occurs before death. Therefore,

$$I_x = \frac{\Pi_x(N_x - N_x P_x)(1 - q_x) + \Pi_x(N_x - N_x P_x) q_x f_x}{N_x - N_x P_x - \left( (N_x - N_x P_x) q_x \right)}$$

(2)

We make two alternate assumptions regarding $f_x$.

A. This assumption is that the times until death and until developing the disease are equal, so that $f_x = q_x$. Under this assumption, $I_x$ reduces to $\Pi_x$, and the incidence rate can be estimated from equation 1. However, as mentioned earlier, there are conceptual differences between $I_x$ and $\Pi_x$.

B. The following alternative assumptions regarding $f_x$ preserve the distinction between $\Pi_x$ and $I_x$. If disease-free times and survival times are assumed to be exponentially distributed, the instantaneous disease risk, $\mu_{C_x}$, and the instantaneous death risk, $\mu_{D_x}$, are constant throughout the interval. Then (see Appendix 1)

$$\Pi_x q_x f_x = \frac{\ln(1 - q_x)}{\ln(1 - q_x) + \ln(1 - \Pi_x)} (1 - q_x)(1 - \Pi_x)$$

(3)

$$I_x = \frac{\ln(1 - \Pi_x)}{\ln(1 - q_x) + \ln(1 - \Pi_x)} \left\{ \frac{1 - (1 - \Pi_x)(1 - q_x)}{1 - q_x} \right\}$$

(4)
ESTIMATING INCIDENCE FROM AGE-SPECIFIC PREVALENCE

Under these formulations, incidence can be estimated either from age-specific prevalence alone (estimate A, expression 1) or from age-specific prevalence and mortality (estimate B, expression 4).

Numerical estimates of glaucoma incidence

The previous section presented the theoretical basis for our method. We now proceed to apply the method to the estimation of age-specific OAG incidence from available data on prevalence and mortality for Wales and Framingham. Because the prevalences in table 1 represent point prevalences at the midpoints of 5-year age groups, the incidence estimated directly from these values by our method would be for age intervals with non-standard limits. For example, from the point prevalences at 42.5 years and 47.5 years, incidence would be estimated for the interval 42.5–47.5 years. By fitting a curve to the observed prevalences, we can obtain point prevalence estimates for any age and thus estimate incidence for standard age intervals. As an example, the incidence for the standard age interval 40–44 years could be calculated from point prevalence estimates at ages 40 and 45 years. This is the approach we followed.

Glaucoma incidence is not constant across age groups, but increases with age (6). Therefore, fitting a linear model to the observed prevalences is less appropriate than, say, fitting a sigmoidal curve such as the logistic (11). We fitted a logistic curve by maximum likelihood (12) and estimated prevalence, \( P_x \), for the beginning of the age intervals (table 2). The age-specific probability of death, \( q_x \), needed to estimate the incidence, \( I_x \), was obtained from general population life tables (13, 14). Variances for the incidence estimates were derived by the "delta method" as shown in Appendix 2.

Results

The age-specific OAG incidence estimates with standard errors, from the Welsh and Framingham studies are presented in table 3. For each study, estimates A and B are derived from equations 1 and 4, respectively. The A and B estimates, as well as their standard errors, are virtually identical. Thus, the incidence estimated as a simple function of age-specific prevalences was similar to the one derived by taking into account age-specific prevalence and mortality.

There is very close agreement between the estimated incidence in Wales and Framingham. This was to be expected, because the age-specific prevalences are similar. The large standard errors reflect the small number of cases.

Discussion

Our main purpose is to present a method of estimating incidence from prevalence in diseases that are irreversible and non-fatal. An additional purpose is to provide such estimates for open angle glaucoma. Although our example deals only with

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Wales (3)</th>
<th>Framingham (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( q_x )</td>
<td>( P_x )</td>
</tr>
<tr>
<td>40–44</td>
<td>0.0127</td>
<td>0.0007</td>
</tr>
<tr>
<td>45–49</td>
<td>0.0211</td>
<td>0.0011</td>
</tr>
<tr>
<td>50–54</td>
<td>0.0354</td>
<td>0.0017</td>
</tr>
<tr>
<td>55–59</td>
<td>0.0592</td>
<td>0.0028</td>
</tr>
<tr>
<td>60–64</td>
<td>0.0936</td>
<td>0.0045</td>
</tr>
<tr>
<td>65–69</td>
<td>0.1427</td>
<td>0.0072</td>
</tr>
<tr>
<td>70–74</td>
<td>0.2178</td>
<td>0.0116</td>
</tr>
<tr>
<td>75–79</td>
<td>0.3073</td>
<td>0.0166</td>
</tr>
<tr>
<td>80–84</td>
<td>0.4018</td>
<td>0.0255</td>
</tr>
</tbody>
</table>

\( q_x \) = probability of death during interval (13, 14).

\( P_x \) = smoothed prevalence proportion for beginning of age interval, determined from logistic fit: \( P_x = 1/(1 + \exp(-a-b-x)) \)

Wales: \( a = -11.099 \) \( b = 0.096 \)
Framingham: \( a = -9.084 \) \( b = 0.068 \)

advantage of fitting a curve, especially with small numbers of cases, is that it smooths the trend.
**Table 3**

*Estimated age-specific five-year incidence rates (and estimated standard errors*) of open angle glaucoma in Wales, 1966, and Framingham, 1978*

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Wales (3)</th>
<th>Framingham (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>At²</td>
<td>Bt</td>
</tr>
<tr>
<td>40–44</td>
<td>0.0004 (0.0015)</td>
<td>0.0004 (0.0015)</td>
</tr>
<tr>
<td>45–49</td>
<td>0.0006 (0.0019)</td>
<td>0.0006 (0.0019)</td>
</tr>
<tr>
<td>50–54</td>
<td>0.0011 (0.0024)</td>
<td>0.0011 (0.0024)</td>
</tr>
<tr>
<td>55–59</td>
<td>0.0017 (0.0032)</td>
<td>0.0017 (0.0032)</td>
</tr>
<tr>
<td>60–64</td>
<td>0.0027 (0.0045)</td>
<td>0.0027 (0.0045)</td>
</tr>
<tr>
<td>65–69</td>
<td>0.0044 (0.0066)</td>
<td>0.0044 (0.0066)</td>
</tr>
<tr>
<td>70–74</td>
<td>0.0053 (0.0098)</td>
<td>0.0052 (0.0098)</td>
</tr>
<tr>
<td>75–79</td>
<td>0.0073 (0.0147)</td>
<td>0.0073 (0.0146)</td>
</tr>
<tr>
<td>80–84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* See Appendix 2 for method of estimation of standard errors.

†: \( I_x = \Pi_x \) (See text, equation 1.)

‡: \( I_x = f(\Pi_x, q_x) \) (See text, equation 4.)

OAG, the method described here may be applicable to other irreversible diseases for which one may assume that the disease does not influence mortality.

In our example, the method permitted a simple estimation of incidence from prevalence data alone, since the inclusion of age-specific mortality in the equation for the incidence did not affect the numerical estimates. Application of the method to other diseases or to other populations could produce different results. When the probability of disease, \( \Pi_x \), is small and the probability of death, \( q_x \), is large, the occurrence of disease before death would be expected to be less frequent than the occurrence of death before development of disease. Thus, \( f_x \) would be less than \( q_x \). Estimate A, which assumes that \( f_x = q_x \) and that \( I_x = \Pi_x \), would be expected to be an overestimate of \( I_x \) because \( \Pi_x \) does not take into account the competing risk of mortality. If \( \Pi_x \) is much larger than \( q_x \), the reverse would be true, that is, \( f_x \) would be higher than \( q_x \). The estimation of \( I_x \) by \( \Pi_x \) would therefore yield an underestimate of the true incidence rate, since some new cases dying before the end of the interval would not be included in the numerator of \( I_x \). A third situation is also possible, in that \( \Pi_x \) and \( q_x \) are not too different. In this case, the assumption that \( f_x = q_x \) is justified, and \( \Pi_x \) and \( I_x \) would be very similar. Because OAG incidence is small relative to mortality, we expected \( \Pi_x \) to be an overestimate of \( I_x \) in our results. This did not occur, possibly because the difference between \( \Pi_x \) and \( q_x \) was not sufficiently large to affect the results. In circumstances where the exponential distribution may be assumed and a very large discrepancy exists between \( \Pi_x \) and \( q_x \), the use of expression 4 to estimate incidence may give different results from those obtained by expression 1.

The assumptions underlying the method are: life-long duration of the disease, equal mortality risk in persons with and without glaucoma, and a stable disease in a stable population. How well does our example fit the latter two assumptions?

According to the assumption that mortality is the same for persons with and without glaucoma, deaths will decrease the number of persons with glaucoma in the population, but will not affect the prevalent proportion. As is often the case with assumptions underlying epidemiologic models, the foregoing assumption may not entirely fit the facts. The life expectancy of persons with glaucoma has been studied among recipients of aid to the blind in California in 1954 (15), and...
among persons first registered as legally blind in Massachusetts during 1940–1959 (16). Both studies found reduced life expectancy for blind persons with glaucoma under the age of 65 years. For those aged 65 years or more, however, mortality experience was the same as in the general population (16) or in recipients of old age assistance, a group comparable in socioeconomic status to the recipients of aid to the blind (15). Most glaucoma patients are not legally blind, therefore data from the previous studies cannot be generalized. The limited data available on life insurance experience of persons with glaucoma (17) has shown no increased mortality in this group.

To the extent that the mortality of glaucoma patients may exceed mortality in the general population, our model will underestimate the incidence of glaucoma. The degree of underestimation, however, is probably not large. In US life tables, the 5-year probability of death in the 50–65 years age range is small, e.g., 8.8 per cent at age 60 years and 3.8 per cent at age 50 years (18). The effect of even doubling these probabilities for glaucoma patients may be small relative to other potential types of error, such as sampling error or ascertainment errors.

We have no evidence to support the contention that glaucoma is a stable disease and little evidence on whether the composition of the Welsh and Framingham populations remained constant over time with respect to glaucoma risk factors. Population migration over a span of 30 or 40 years could be substantial and the resulting population changes could influence disease incidence and prevalence in a given location. After the first 25 years of the Framingham Heart Study, about one-fourth of the original cohort had migrated (4). On the other hand, a comparison of examinees and non-examinees (mostly emigrants) in the Framingham Eye Study with respect to 15 demographic, medical, and social factors that had been ascertained as part of the Framingham Heart Study, "only age was notable as different, with the examinees being younger" (4). Because we employed age-specific data in the present study, this age difference by itself should not bias our incidence estimates.

The small numbers in our examples subject the estimated incidence rates to a considerable sampling error. Another possible source of error is underascertainment of cases. For example, in a recent publication, Kahn and Milton (19) suggest that the reported prevalence of OAG in the Framingham Eye Study may be too low. It is not possible, however, to validate our estimates, since population-based incidence data are not available.

Because of the sampling and non-sampling errors and the imperfections of the assumptions in the model, the estimates we have prepared should be viewed as nothing more than gross approximations of the true incidence rates, but even these may be helpful for certain purposes, such as the planning of cohort studies. We have used this method to determine sample size requirements for epidemiologic studies of OAG (20).

REFERENCES

APPENDIX 1

Estimating Incidence According to Constant Risk Assumption

Under the exponential assumptions, the probability of surviving until time $t$ ($0 < t < 1$) in interval $x$, is

$$e^{-\mu_D x t},$$

and the probability of remaining disease-free until time $v$ ($0 < v < t$) in interval $x$ is

$$e^{-\mu_G x v}.$$

Then the probability of developing the disease during the interval and then dying before the end of the interval is

$$\prod_x q_x f_x = \int_0^1 \mu_D x e^{-\mu_D x t} \left[ \int_0^t \mu_G x e^{-\mu_G x v} dv \right] dt$$

Evaluation of the integrals yields

$$\prod_x q_x f_x = \frac{\mu_D x e^{-\mu_D x + \mu_G x t}}{\mu_D x + \mu_G x} - e^{-\mu_D x} + \frac{\mu_G x}{\mu_D x + \mu_G x}$$

However, we need to express $\mu_D x$ and $\mu_G x$ in terms of $q_x$ and $\prod_x$, where

$$q_x = \int_0^1 \mu_D x e^{-\mu_D x t} dt = 1 - e^{-\mu_D x}$$

and

$$\prod_x = \int_0^1 \mu_G x e^{-\mu_G x t} dt = 1 - e^{-\mu_G x}.$$
Appendix 2

Estimating the Variance for the Incidence Estimates

Incidence estimate A: \( I_x = \prod_x \).
Incidence estimate B: \( I_x = f(\prod_x, q_x) \).

The variance is estimated by the method shown in Armitage (11), which is often referred to as the "delta method."

First, for the more complicated estimate B,

\[
\text{Var}(I_x) = \left( \frac{\delta I_x}{\delta \prod_x} \right)^2 \text{Var}(\prod_x) + 2 \left( \frac{\delta I_x}{\delta \prod_x} \right) \text{Cov}(\prod_x, q_x) + \left( \frac{\delta I_x}{\delta q_x} \right)^2 \text{Var}(q_x).
\]

If \( q_x \) is assumed to be a constant, then the second and third terms of the above equation are zero. Therefore,

\[
\text{Var}(I_x) = \left( \frac{dI_x}{d\prod_x} \right)^2 \text{Var}(\prod_x).
\] (2-A)

That this is also the variance estimate for incidence estimate A is readily seen. By the delta method,

\[
\text{Var}(\prod_x) = \frac{1}{N_x} \left( \frac{P_x + 1}{1 - P_x} \right)^2 \text{Var}(P_x) = \left( \frac{1}{1 - P_x} \right)^2 \text{Var}(P_x) = \left( \frac{1}{1 - q_x} \right)^2 \text{Var}(P_x) + \left( \frac{1}{1 - q_x} \right) \text{Var}(P_x + 1).
\] (2-B)

Assuming that the prevalence counts, \( N_xP_x \), are Poisson random variables,

\[
E(N_xP_x) = \text{Var}(N_xP_x),
\]

so that

\[
\text{Var}(P_x) = P_x/N_x,
\]

and

\[
\text{Var}(P_x + 1) = P_x + 1/N_x.
\]

Therefore, from (2-A) and (2-B),

\[
\text{Var}(I_x) \approx \left( \frac{dI_x}{d\prod_x} \right)^2 \left( \frac{P_x + 1}{1 - P_x} \right)^2 \frac{P_x}{N_x} + \left( \frac{1}{1 - P_x} \right)^2 \frac{P_x + 1}{N_x + 1}.
\] (2-C)

Values of \( P_x \) and \( P_x + 1 \) were obtained from table 2. Values of \( N_x \) and \( N_x + 1 \) were obtained from table 1 by dividing the mid-interval population by \( 1 - q_x \) to estimate the population at the beginning of the interval. Values of \( dI_x/d\prod_x \) were determined as follows:

Estimate A: \( I_x = \prod_x \),

\[
dI_x/d\prod_x = 1.
\]

Estimate B: \( I_x \) estimated by the constant risk assumption (see equation 4)

\[
dI_x/d\prod_x = \left\{ \frac{\ln(1 - \prod_x)}{\ln(1 - \prod_x) + \ln(1 - q_x)} \right\} d/\prod_x \left\{ \frac{1 - (1 - \prod_x)(1 - q_x)}{1 - q_x} \right\}
\]

\[
+ \left\{ \frac{\ln(1 - \prod_x)}{\ln(1 - \prod_x) + \ln(1 - q_x)} \right\} \left\{ \frac{1 - (1 - \prod_x)(1 - q_x)}{1 - q_x} \right\}
\]

\[
= \frac{(1 - \prod_x)(1 - q_x)[\ln(1 - \prod_x)][\ln(1 - \prod_x) + \ln(1 - q_x)] - [1 - (1 - \prod_x)(1 - q_x)] \ln(1 - q_x)}{(1 - q_x)(1 - \prod_x)[\ln(1 - \prod_x) + \ln(1 - q_x)]^2}.
\]