Medical Statistics: an Example from Diabetes Epidemiology

Niels Keiding
University of Copenhagen, Copenhagen (Denmark)

The purpose of this contribution is to introduce some concepts and methods of modern survival analysis such as those used in chronic disease epidemiology. This will be done through a concrete example, concerning the estimation of the incidence of diabetes and the death rate of diabetics based on a prevalent sample from Fyn county, Denmark.

Fyn county in Denmark contains about 450,000 inhabitants. On July 1, 1973, 1499 of them suffered from insulin-dependent diabetes mellitus ("diabetes" for short). This was ascertained by recording all insulin prescriptions in the National Health Service files for this county during a five month period covering the above date, and subsequently checking each patient’s medical record at the general practitioner and, when relevant, the hospital. The data was collected by Green et al. (1981).

Two different problems were studied: estimating the age-specific mortality of diabetics (number of deaths) based on (prospective) follow-up of this prevalent population, and a retrospective estimation of diabetes incidence in the years before 1973, based on the information in the sample on (calendar time, age) at onset and additional historical information on diabetes mortality and population censuses. Our model is a non-homogeneous illness-death (Markov or semi-Markov) process for each person with intensities as illustrated in the diagram

\[
\begin{align*}
\text{healthy} & \xrightarrow{\alpha(a)} \text{diabetes} \\
\mu(a) & \xleftarrow{\nu(a,d)} \text{death}
\end{align*}
\]

where the incidence \(\alpha(a)\) and the mortality \(\mu(a)\) of the healthy (i.e., non-diabetic) depend on age \(a\), whereas the mortality \(\nu(a,d)\) of the diabetic (the death rate) may in addition depend on duration \(d\), and all intensities may depend on further covariates, in particular sex and calendar time. A general discussion of the application of this framework for formalizing epidemiological concepts such as age-specific incidence and prevalence was given by Keiding (1991).

For the prospective study of the death-rate, the survival status for all patients was assessed by January 1, 1982, using the Danish centralized person registry. The basic time variable was age. Additional time variables are duration of disease and
calendar time. A special problem is that patients do not all enter the study at the same age.

Because of this problem of delayed entry (patients only being at risk for dying of diabetes after they have acquired the disease), we focus on the conditional survival distribution, given survival until a certain age, for which we use the standard non-parametric maximum likelihood estimator, the Kaplan–Meier estimator with delayed entry (left truncation). Additional covariates may be included in a regression model for survival data; this is illustrated by incorporating disease duration in a Cox regression model (with time- (i.e., age-) dependent covariate). A survey of problems in studying delayed entry was given by Keiding (1992).

The death-rate in a seven-year follow up period was evaluated by Green and Hougaard (1984). An extended discussion of the analysis of this data was given by Andersen et al. (1992).

The second problem regarded the possibility of exploiting the retrospective information in the prevalent sample of July 1, 1973 on (calendar time, age) at onset for each patient to assess diabetes incidence \( \alpha(t, a) \), \( 1933 \leq t \leq 1973, 0 \leq a \leq 31 \) years.

Obviously, the prevalent sample of July 1, 1973 represents a selected sample of patients, containing only the survivors. Heuristically, each surviving patient represents not only his/herself but also the “ghosts” (Turnbull, 1976) who died before sampling. The expected number of patients necessary to generate one survivor is \( p^{-1} \) where \( p = \) this patient’s survival probability from onset to July 1, 1973. To estimate \( p \) (for \( 0 \leq a \leq 30 \) years) we may apply the analysis by Ramlau-Hansen et al. (1987) who fitted a Cox regression model for the relative mortality (Andersen et al., 1985) to historical hospital data on diabetes death-rate. Data on population censuses is available in official Danish vital statistics.

A smooth estimate of diabetes incidence \( \alpha(t, a) \) in the Lexis diagram (the calendar time \( x \) age plane) may now be obtained using bivariate kernel smoothing (Keiding et al. 1989, Keiding 1990, McKeague and Utikal 1991). The basic trends show increase with calendar time for a given age (slightly more for males than for females) and increase with age for given calendar time, with an indication of a peak at about the time of puberty. Viewed on a cohort basis, consistent increase of diabetic incidence with age are observed in the later cohorts, but this pattern is not very obvious in the earlier cohorts.

References


Niels Keiding
Statistical Research Unit
Faculty of Medicine
University of Copenhagen
Copenhagen, Denmark