10-1-2005

Estimation and Projection of Indicence and Prevalence Based on Doubly Truncated Data with Application to Pharmacoepidemiological Databases

Henrik Stovring
University of Southern Denmark, Research Unit of General Practice, stovring@biostat.au.dk

Mei-Cheng Wang
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mcwang@jhsph.edu

Estimation and projection of incidence and prevalence based on doubly truncated data with application to pharmacoepidemiological databases

Henrik Støvring and Mei-Cheng Wang

October 7, 2005

Abstract

Incidence of diseases are of primary interest in any epidemiological analysis of disease spread in general populations. Ordinary estimates obtained from follow-up of an initially non-diseased cohort are costly, and so such estimates are not routinely available. In contrast, routine registers exist for many diseases with data on all detected cases within a given calendar time period, but lacking information on non-diseased. In the present work we show how this type of data supplemented with data on the past birth process can be analyzed to yield age specific incidence estimates as well as lifetime prevalence. A non-parametric model is studied with emphasis on the required assumptions, and a brief outlook on the analysis of the non-stationary case with calendar trends in age-specific incidence is given. The developed methods are applied to case cohort data on treatment with antidiabetic medications and projections are provided for both diabetes incidence and prevalence. As projection of diabetes prevalence requires estimation of the distribution of disease durations, two novel approaches for this estimation is studied, a parametric and a non-parametric, respectively.

1 INTRODUCTION

Chronic diseases with long durations are at the center stage of public health interest in western, modern societies, since they affect many and induce high costs for patients as well as society. The costs for patients include higher comorbidity, loss of life-years as well as quality of life, and for society loss of productive life years and expenses incurred by provision of health care services. A proper understanding of the disease and its costs requires knowledge of the fundamental epidemiological measures incidence, prevalence, and mortality. These are however typically difficult and costly to measure, and so alternatives based on either inexpensive or existing data would be of high interest.

One such datatype is the so-called *case cohort data*, which consists of all occurences of new cases, be it disease or treatment onset, within a fixed time window. Case cohort designs are generally considered to be efficient, in particular for diseases with a low rate of occurrence; see (Mantel 1973), (Prentice 1978), (Oakes 1981), (Thomas 1981), (Lubin and Gail 1984), and references therein. A case-control design typically collects data from all the cases and selects one or more time-matched 'controls' to match each case to conduct relevant analysis. In this paper we consider a case cohort which consists of only the cases—in our case identified by means of a pharmaco-epidemiological database—possibly supplemented with additional information on the process of initiating events, for example the birth process of the general population.

In case cohort studies it is common to collect case data within a calendar time period. In this paper we assume that the sample includes all subjects who have advanced to a certain end-point (failure event) within a given calendar time period, and that the time origin (initiating event, birth time) of each case can be retrospectively identified. So far statistical methods for this type of data have not (to the extent of the authors' knowledge) been extensively studied, when the rate of initiating events is not assumed constant over calendar time. In this paper we introduce such a methodology which yields a non-parametric maximum likelihood estimate of the age-specific incidence distribution based solely on case cohort data,

and allows supplementing with a known birth rate. The non-parametric method does not directly provide measures of the uncertainty of the estimate, and so we propose a bootstrap method for obtaining measures of this uncertainty.

Note that the case data considered in this paper provide information which is different from the information of the cases in the cohort case-control studies mentioned earlier (Mantel 1973; Prentice 1978; Oakes 1981; Thomas 1981; Lubin and Gail 1984), although the two types of data do share common characteristics. As pointed out in (Prentice 1986, p4), in cohort case-control studies the failure time is usually defined as the time from the beginning of follow-up to a failure event, which is different from the failure time considered in the current setting. The failure times from the cases in the former studies can be thought of as right-truncated data, in contrast with the doubly-truncated data investigated in this paper.

Further, it is important to realize that we cannot supplement with retrospective information on age at incidence among prevalents, see (Keiding, Holst, and Green 1989). This is due to the fact that onset ages are not observable outside the observation window in pharmacoepidemiological databases. As another consequence, the "usual" analysis of disease duration based on delayed entry is not possible here, as we cannot condition on time since onset prior to the start of the observation window.

Before presenting the theory for estimation and projection in Section 3, we first introduce the data in Section 2 which is used for the application. In Section 4 we present analyses of incidence of use of anti-diabetic medications and present projections of both incidence and prevalence based upon these.

2 DATA

2.1 Case cohort data on antidiabetic treatment

For the period 1992–2003 the Odense Pharmaco-epidemiological Database (OPED) contains subject specific information on all prescriptions for subsidized medications redeemed at a pharmacy in the County of Fyn, as well as information on births, deaths and migration into and out of the County of Fyn. For each individual we identified all prescriptions of antidiabetic agents in OPED. The antidiabetic drugs are characterized by the first three characters being A10 (The WHO Collaborating Centre for Drug Statistics Methodology 2001). We will not distinguish between the various types of antidiabetic treatments, such as for example insulin (A10A) and oral antidiabetics (A10B). Incident events are defined to be the first treatment event observed in the time window for subjects who did not have any previous events during a one year run-in period. The run-in period was either started at the start of the database or at the time of first immigration into Fyn of the subject, if the subject immigrated into Fyn during the observation period. Note, that this may well introduce a calendar-time-dependent misclassification and hence bias, cf. (Støvring, Andersen, Beck-Nielsen, Green, and Vach 2003), but this will be ignored in the following as we are not studying secular trends in incidence. Also note, that by definition, these data will only allow us to study incidence and prevalence of pharmacologically treated diabetes. We will thus use the words "treated" and "diseased" interchangeably, and ask the reader to keep in mind that the present analysis only pertains to pharmacologically treated diabetes.

2.2 Birth rates

For the period 1891-2003, available data from Statistics Denmark were used to determine annual, national birth counts for each gender. To estimate the number of births within the county of Fyn, data was obtained on population size for Denmark as a whole, as well as for Fyn with the objective of rescaling. Population counts were available roughly at five year intervals (1901, 1906,..., 1921, 1925, 1930, ..., 1970, 1976, 1981, 1986, 1990, 1995, 1998, 1999,...,2003) for Fyn, whereas nationwide data was available annually from 1970 and onward, and otherwise similar to those for Fyn given above.

To estimate the number of births in the county of Fyn, we scaled national birthrates by the relative population size in the county of Fyn compared to the total population of Denmark. The underlying assumption is that the fertility rate on Fyn is similar to national rates, which seems plausible given the small size of Denmark and the relatively homogeneous composition of the population. As population counts are not

available annually we interpolated the population data based on simple piecewise linear regression, with cut points at 1920, 1970, and 1996, cf. Figure 1.

[Figure 1 about here.]

Overall, Fyn hold 9%-10% of the Danish population during most of the twentieth century and the fit seems very good. The sudden drop in 1920 is due to the reunion of North Slesvig with Denmark after having been part of Germany from 1864.

In subsequent analyses the missing proportions were replaced with the predicted, while the observed proportions were retained. When we combined this with the national birth rates, we could compute the number of births in the county of Fyn as the product of the number of births in Denmark and the proportion of the Danish population living in the County of Fyn. Since no observations were available for the ten year period 1891-1900, we predicted the annual number of births in this period from a linear extrapolation of the birth counts in the period 1901-1910. The resulting gender specific annual birth rates in the County of Fyns are presented in Figure 2. The rates appear clearly non-stationary, highlighting the need for methods that account for this.

[Figure 2 about here.]

Note, that all estimated number of births are treated as fixed in subsequent analyses.

3 THEORY AND SET-UP

Before proceeding, we want to point out that it is well understood that in the analysis of right-truncated data the presence of truncation would result in biased inferences for the time to event outcome variable. As an interesting contrast, under certain conditions, the bias caused by double truncation could be removed because of the presence of both left and right truncation. This phenomenon will later be explained by an expression of the marginal density for the observed events.

We now introduce the notation used in the paper. Let U be the calendar time of the initiating events (births). Let Y be age at onset if the disease occurs before death, and ∞ in the absence of disease before death. Let the population density function (pdf) of Y be $f(y|u)$, and the associated cumulative distribution function (cdf) $F(y|u)$. Further, let Z_0 be age at death if $Z_0 < Y$, that is disease does not occur before death. If $Y > Z_0$, we let $Y = \infty$, and otherwise we let $Z_0 = \infty$. Let R be duration of disease with cdf $K(r|y,u)$, where R is only defined if $Y \leq Z_0$. Define Z_1 to be age at death if $Y \leq Z_0$, undefined otherwise. For ease of reading we will at times denote F as F_Y, K as K_R and so on.

Since not all subjects will experience disease prior to death, the pdf of Y, $f(y|u)$, is a mixture distribution with two components:

$$f(y|u) = \pi_\infty(u) f^*(y|u) + (1 - \pi_\infty(u)) \tag{1}$$

where $\pi_\infty(u)$ is defined as $P(Y < \infty|u)$, i.e. it is the probability of disease occuring before death, and $f^*(y|u)$ is the conditional pdf of Y given that $Y < \infty$, i.e. $Y \leq Z_0$. Note, that since $\pi_\infty(u)$ is the probability of disease occurring before death, it is the lifetime prevalence.

Assume that we observe all events of onset, Y, occurring within the calendar time observation window $[0; \tau_0)$. Assume that the occurrence of births follows a Poisson process with intensity $\phi(u)$ for $u \leq \tau_0$, and that $y^+ = \sup\{y : F^*(y|u) < 1\}$ exists and is finite for all $u \leq \tau_0$, where y^+ is the maximal observable age at onset before death. We can then normalize the birth intensity $\phi(u)$ to a density g on $[-y^+; \tau_0)$:

$$g(u) = \frac{\phi(u)}{\int_{-y^+}^{\tau_0} \phi(s)\, ds} \tag{2}$$

with associated cumulative distribution function G. In principle ϕ could well depend on covariates, but since we consider ϕ either known or constant, we will ignore this.

We will in the following assume $(U_1, Y_1), \ldots, (U_n, Y_n)$ to be independent and identically distributed (iid).

A crucial assumption to consider is whether or not we have calendar time stationarity with respect to age of onset, i.e., whether or not the following assumption is valid:

(S1) The age of onset is independent of time of birth, i.e., $F(y|u) = F(y)$.

Although we here have knowledge of the birth process, this will not be the case in many applications. Hence we also consider the situation with calendar time stationarity of the birth process:

(S2) Assume that the occurence of initiating events started in the distant past and that the rate of occurence has been stabilized. Or, quantitatively, assume that $u_x = \inf\{u : \phi(u) > 0\}$ is small enough so that $u_x \leq -y^x$, and that G_x is uniform on $[-y^x; \tau_0)$.

3.1 Stationary incidence, known birth process

When (S1) holds, the joint density of the observed (u, y) can be written as follows:

$$\begin{aligned} p(u, y| -U \leq Y \leq \tau_0 - U) &= \left[\frac{g(u)I(-y \leq u \leq \tau_0 - y)}{G(\tau_0 - y) - G(-y)}\right] \\ &\quad \times \left[\frac{\{G(\tau_0 - y) - G(-y)\}f^*(y)I(y \leq y^+)}{\int_0^{y^+}\{G(\tau_0 - s) - G(-s)\}f^*(s)ds}\right] \quad (3) \\ &= p_c(u|y)p_m(y) \quad (4) \end{aligned}$$

where $p_c(u|y)$ can be interpreted as the density of birth times conditional on y being observed in $[0; \tau_0)$, and $p_m(y)$ is the marginal density for the observed y weighted with $w_i = G(\tau_0 - y) - G(-y)$, i.e., the probability of birth occuring within the interval $[-y; \tau_0 - y)$.

When g is known, then so is p_c, as are the weights in p_m. It is thus straightforward to compute the maximum likelihood estimate of F^* based on the weighted observations:

$$\widehat{F}^*(y) = \frac{\sum_{i:y_i \leq y} w_i^{-1}}{\sum_{i=1}^n w_i^{-1}} \quad (5)$$

The estimate thus places mass $w_j^{-1} / \sum w_j^{-1}$ at each jump point j, where j corresponds to the observation number in the ordered set of Y_i. If all weights are equal, the above formula corresponds to the ordinary formula for non-parametric estimation of a cdf in the uncensored case, putting mass n^{-1} at each jump point.

With the estimate of the conditional cdf F^* it is possible to obtain an estimator of the unconditional F utilizing their relationship given in Equation (1). What we need is an estimate of π_∞, which may be obtained from noting that the occurrence rate of incident events I^{tr} at any calendar time point is given by

$$\begin{aligned} I^{\text{tr}}(t) &= \int_{-\infty}^{t} \phi(u)f(t-u)I(t-u \leq y^+)du \quad (6) \\ &= \pi_\infty \int_{-\infty}^{t} \phi(u)f^*(t-u)du \quad (7) \end{aligned}$$

where the indicator function $I(t-u \leq y^+)$ is needed, since the occurrence rate obviously does not include those for which onset never happens, that is when $y = t - u > y^+$ or equivalently that $y = t - u = \infty$. Integrating this over the observation window, we find

$$\begin{aligned} \int_0^{\tau_0} I^{\text{tr}}(t)dt &= \pi_\infty \int_0^{\tau_0}\int_{-\infty}^{t} \phi(u)f^*(t-u)du\,dt \quad (8) \\ &= \pi_\infty \int_{-\infty}^{\tau_0} \phi(u)\left\{\int_{\max(u,0)}^{\tau_0} f^*(t-u)dt\right\}du \quad (9) \\ &= \pi_\infty \int_{-\infty}^{\tau_0} \phi(u)\left\{\int_{\max(0,\,-u)}^{\tau_0-u} f^*(t)dt\right\}du \quad (10) \\ &= \pi_\infty \int_{-\infty}^{\tau_0} \phi(u)\left\{F^*(\tau_0 - u) - F^*(\max(0,\,-u))\right\}du \quad (11) \end{aligned}$$

from which it follows that

$$\pi_\infty = \int_0^{\tau_0} I^{\text{tr}}(t)dt \Big/ \left[\int_{-\infty}^{\tau_0} \phi(u)\left\{F^*(\tau_0 - u) - F^*(\max(0,\ -u))\right\} du\right] \tag{12}$$

Plugging in the MLE of F^*, we obtain an estimate of π_∞, since ϕ is known and $\int_0^{\tau_0} I^{\text{tr}}(t)dt$ is estimated by the total number of observed incidences over the interval $[0; \tau_0)$. Having obtained the MLE of F^* together with an estimate of π_∞, we can use Equation (1) to compute an estimate of F, the unconditional cdf of age at onset.

3.2 Stationary incidence and stationary birth process

When both (S1) and (S2) holds, the marginal density of the observed y's can be further simplified

$$p(y) = \left[\frac{\{\tau_0/(\tau_0 + y^x)\}f^*(y)I(y \le y^+)}{\tau_0/(\tau_0 + y^x)}\right] = f^*(y)I(y \le y^+) \tag{13}$$

Here the density function of the observed y's coincides with the population density function f^* of the observable onset times, Y. In the case when only age at onset distribution is of interest, and not lifetime prevalence, the 'usual methods' are thus applicable to the case data to estimate f^* by putting equal weights on all observations as noted above.

If, however, we are also interested in the unconditional density, i.e. $f(y)$, we need an estimate of π_∞ to be able to proceed. Above, this was obtained from our knowledge of the birth process, and in principle we could exploit this again. However, in situations where a stationary birth process is assumed, this is typically because we lack information on the birth process. Thus it may in such situations be necessary with alternative approaches. One obvious way to proceed is the following: In the time window where information is collected on incident cases, we also collect information on deaths—either for all or a random sample—and classify them according to whether or not they had experienced disease. The relative frequency of diseased deaths will then be an estimate of π_∞ under stationarity assumptions with respect to the birth process, the incidence process, and the mortality. With this estimate of π_∞ we may then estimate the unconditional F.

3.3 Non-stationary incidence

When (S1) does not hold—or rather when we are not willing to make this assumption—the likelihood becomes substantially more complicated. In principle this can be handled by introducing a parameter vector θ relating the incidence density to the time of birth.

We can in principle still undertake the rewriting presented in Equation (4), with the modification that the density term $p_m(y)$ now depends on parameters θ, i.e.

$$p_m(y|\theta) = \frac{\{G(\tau_0 - y|\theta) - G(-y|\theta)\}f^*(y|\theta)}{\int_0^{y^+}\{G(\tau_0 - s|\theta) - G(-s|\theta)\}f^*(s|\theta)ds} \tag{14}$$

This density does unfortunately not directly permit use of the approach presented above for finding a non-parametric estimate of $f^*(y|\theta)$, nor for finding the corresponding estimate of $\pi_\infty(\theta)$.

One alternative is to set up a full likelihood by considering a full parametric model of both age of onset and age of death, but we will not go into further details here and instead commend this as a topic for future research.

3.4 Projection of incidence and prevalence

Projection of incidence is possible both inside and outside the observation window by application of the formula in Equation (6), when the birth process is known and incidence is assumed stationary. In the application studied here, the birth process is known for $u \le \tau_0$. For $u > \tau_0$ it must be projected. As a simple starting point, we will carry the last observed value of the birth process forward, i.e. let $\phi(u) = \phi(\tau_0)$

for $u > \tau_0$. As the incidence density $f(y|\theta)$ is estimated, we can plug in this estimate to obtain a projection of incidence.

Projection of prevalence at time t is in general possible using one of two approaches. The first is based on the incidence rate and distribution of durations:

$$P^{\text{tr}}(t) = \int_{-\infty}^{t} I^{\text{tr}}(w)(1 - K(t-w|w))dw \tag{15}$$

where $K(t-w|w) = P(R \leq t-w|U+Y = w)$. We propose two strategies for estimating K, a parametric and a non-parametric.

The second approach is based on the birth rate and the probability of being prevalent and alive at time t

$$P^{\text{tr}}(t) = \int_{-\infty}^{t} \phi(u)P(Y \leq t-u, Z_1 > t-u)du \tag{16}$$

Note that $P(Y \leq t-u, Z_1 > t-u)$ is age-specific prevalence. We will not further discuss this approach since age-specific prevalence is not readily estimable for all birth cohorts in the present setting, in particular not when allowed to depend on birth time.

As above, when estimating and projecting incidence, there are two natural assumptions to consider with respect to the stationarity of mortality among diseased. The first is:

(M1) Age at death among diseased is independent of birth, i.e., $H(z|u, Y = y) = H(z|Y = y) \equiv H_y(z)$.

Note, that H_y is explicitly allowed to depend on age at onset.

The second assumption states that the occurrence rate of incident cases is constant with respect to calendar time, and is thus identical to jointly assuming (S1) and (S2). For convenience we will label the joint assumption of (S1) and (S2) as (M2).

We first focus on projection of prevalence at t, based on a parametric estimate of disease duration.

Assume that both (S1) and (M1) are satisfied. This implies that the distribution of disease duration is stationary with respect to calendar time, i.e. R is independent of calendar time of onset and so $K_R(r|w) = K_R(r)$. A parametric estimate of disease duration may then be based on a likelihood composed of two types of contributions:

1. For the sub-cohort of incidence cases (i.e., diabetic incidences occurring within the window $[0; \tau_0)$), their likelihood contribution is formed by the usual right-censored data, i.e.

$$\prod_i k_R(r_i)^{\delta_i}(1 - K_R(r_i^+))^{1-\delta_i} \tag{17}$$

   where r_i is observed duration, r_i^+ is a right censored duration, δ_i is an indicator of censoring (1 if uncensored, 0 if censored), and k_R is the pdf associated with K_R, the distribution of durations.

2. For the sub-cohort of subjects who are prevalent at time 0 and dying within the time window $[0; \tau_0)$, the likelihood contribution is

$$\prod_i \frac{\int_0^{-z^+} I^{\text{tr}}(w)k_R(w + s_i)dw}{C_0} \tag{18}$$

   where C_0 is a standardizing constant $= \int_0^{\tau_0}\int_0^{-z^+} I^{\text{tr}}(w)k_R(w + v)dw\,dv$. Essentially the idea is to use the weighted pdf similar to (14) but replace the birth intensity by the diabetic incidence intensity, and integrate out the unobserved truncation time (time from incidence to calendar time 0, w), with respect to the distribution of W. We do not include contributions which are both left and right censored as they are subject to length biased sampling, which would substantially complicate estimation while only contributing little additional information.

As the observation window is relatively short in comparison to most durations, only a parametric k_R can be estimated, as the likelihood does not otherwise control the tail probability of k_R.

The non-parametric approach is best described by setting up an Illness-Death model as depicted in Figure 3. Again we assume (S1) and (M1).

[Figure 3 about here.]

First, let both λ and α_H depend on current age, y, and let α_I depend on both age at onset, y, and current age, $y+r$. The pdf of durations, k, is given by:

$$\begin{aligned} k(r) &= P(Z_1 - Y \in [r; r+\delta r) | Y \leq Z_0) && (19)\\ &= \frac{P(Z_1 - Y \in [r; r+\delta r), Y \leq Z_0)}{\pi_\infty} && (20)\\ &= (\pi_\infty)^{-1} \int_0^{y^+} P(Z_1 - y \in [r; r+\delta r), Y \in [y; y+\delta y), Y \leq Z_0) dy && (21)\\ &= (\pi_\infty)^{-1} \int_0^{y^+} \lambda(y) \exp\left[- \int_0^y \lambda(s) + \alpha_H(s) ds \right] \\ &\qquad \times \alpha_I(y, y+r) \exp\left[- \int_y^{y+r} \alpha_I(y,s) ds \right] dy && (22) \end{aligned}$$

Let us now assume that $\alpha_I(y, y+r) = \alpha_I(y+r)$ for all y, i.e. the mortality rate among diseased does not depend on age at onset. We then get the following rewriting:

$$\begin{aligned} k(r) &= (\pi_\infty)^{-1} \int_0^{y^+} f_Y(y) \frac{h^*_{Z_1}(y+r)}{S^*_{Z_1}(y)} dy && (23)\\ &= \int_0^{y^+} f^*_Y(y) \frac{h^*_{Z_1}(y+r)}{S^*_{Z_1}(y)} dy && (24) \end{aligned}$$

where $S^*_{Z_1}(a) = \exp[-\int_0^a \alpha_I(s) ds]$ and $h^*_{Z_1}(a) = \alpha_I(a) \exp[-\int_0^a \alpha_I(s) ds]$. Note, that this pdf and survivor function correspond to the distribution that may be estimated, i.e. $P(Z_1 \in [a; a+\delta a) | Y \leq a)$, which is the same as $P(Z_1 \in [a; a+\delta a) | Y \leq a, Y \leq Z_0)$.

In the paper we have so far developed methodology for estimation of F^*_Y. It is now interesting that the very same methodology may be used to obtain an estimate of $H_{Z_1|Y<a}(a)$. All that is needed is to identify all diseased subjects who die within the observation window and record their age at death (or a sample thereof), and then use these and the weights based on the birth process to construct a non-parametric estimate of $H_{Z_1|Y \leq a}(a)$.

If, finally, complete stationarity is assumed, i.e. (M1) and (M2), we suggest taking advantage of the following well-known epidemiologic formula (see for example (Keiding 1991))

$$P^{\text{tr}} = I^{\text{tr}} \mu_R \tag{25}$$

where $\mu_R = E(R)$ is the mean duration of disease. This mean can readily be estimated, since $E(R) = E(Z_1 | Y \leq Z_0) - E(Y | Y \leq Z_0)$, and estimates of the latter two means can be obtained from the estimates of F^*_Y and $H_{Z_1|Y \leq Z_0}$, respectively.

4 ANALYSIS OF ANTIDIABETIC TREATMENT

Tables 1 and 2 gives basic descriptive statistics of the studied population. Table 1 shows the number of incidence events tabulated by gender, birth period and calendar year, which is used for estimating age-specific incidence. As above, all analyses are based on using a one-year run-in period. Table 2 shows number of observed durations (onset and death observed), right censored durations (onset observed, no death observed), and number of doubly truncated durations (prevalent at 0, death in $[0; \tau_0)$).

[Table 1 about here.]

[Table 2 about here.]

4.1 Complete stationarity

Although the birth process is known in our setting, we for comparison present an analysis based on assuming stationarity for the birth process, the incidence process, as well as the mortality process among diseased. We first classified all deaths according to whether or not a previous redemption of antidiabetics had been observed, considering all with such a redemption to be diabetics. The lifetime prevalence, π_∞, was for females estimated at 9.68% (95% Confidence Interval: 9.35%; 10.02%) and for males at 10.86% (10.51%; 11.22%), where both confidence intervals are binomial exact. The estimated incidence distribution, F, stratified on gender is shown in Figure 4.

[Figure 4 about here.]

4.2 Stationarity of incidence, known birth process

If we assume stationarity of the incidence distribution only, we make a non-parametric analysis based on the weighted likelihood given in Equation (4) and the estimator of π_∞ in Equation (12). With the gender specific birth rates, we estimated gender specific estimates of F, π_∞, and hence F, from the observed events and associated ages at the events. The resulting estimates of the incidence distribution F is displayed in Figure 5.

[Figure 5 about here.]

We see that the incidence distribution for both genders are made up of two components: The first component is a more or less constant density for ages below 40 years (the linear part in F), whereas the second is a much higher, unimodal density for ages above 40 years which vanishes for ages above 80 (the sigmoid shaped part of F). For males the estimated lifetime prevalence, $\widehat{\pi}_\infty$, is 15.61% (15.58%; 15.65%), whereas for females it is 13.77% (13.74%; 13.81%). Both confidence intervals are computed using bootstrap with a thousand replications. The confidence intervals are very narrow which reflects the high statistical efficiency of the weighted likelihood approach—which in turn partly comes from the strong assumption of stationarity.

The shape of F is quite similar to the unweighted estimate, whereas the estimated lifetime prevalences are substantially higher than those estimated above. The major explanation is of course lack of stationarity of the true lifetime risk and/or the disease duration: The estimate of π_∞ based on disease status among observed deaths takes most of its information from the older cohorts as they are the ones with high mortality. If the older cohorts had lower lifetime prevalences and/or previously had relatively higher mortality among diseased compared to non-diseased, this will result in a decreased estimate of π_∞. This would also be attenuated if older cohorts are larger than younger cohorts, as is indeed the case here, cf. Figure 2. Contrastingly, when indirectly estimating π_∞ based on weighting with the birth process, the estimate can be viewed as a weighted average of π_∞ over the entire interval for the birth process $[-y^x; \tau_0 - y)$.

4.3 Projection of diabetes incidence and prevalence

In the completely stationary situation, where (S1) and (S2) are both assumed to hold, the projected annual incidence is a constant number equaling the lifetime prevalence times the annual number of births. As the annual number of births are usually not observable in such settings, an alternative is needed. In the spirit of estimating π_∞ from the treatment status among deaths, one could take the total annual number of deaths as an estimate of the number of births. If the population is in a completely stationary state, the annual number of deaths must on average equal the average annual number of births. In our setting the observed numbers of deaths over the 11 year period are 29,871 for females and 29,816 for males yielding projected, annual incidences of 262.8 for females and 294.3 for males. Prevalence projection based on estimated mean disease duration yields 4111.9 for females and 3888.5 for males.

In Figure 6 the incidence is projected based on the weighted, non-parametric estimate of F obtained above in Section 4.2, i.e. with known birth intensity and stationary incidence. All annual birth counts after 2003 are set to the number of births observed in 2003. A projection twenty years into the future is likely to be very inaccurate, and so only short term predictions (five to ten years) should be considered trustworthy.

Also note that the observed incidence strongly suggests a departure from non-stationarity, and so actual incidences are likely to be higher than those projected from a stationarity assumption.

[Figure 6 about here.]

The projected incidences show a small but persistent decline for 2004-2025. The general level is much higher than above, reflecting the higher estimate of π_∞ obtained from using the known birth distribution, but correspond well with observed incidences.

For projection of prevalence, we apply both the parametric and non-parametric estimation of K developed above. For the parametric estimation we used a Weibull distribution with log-transformed parameters, i.e.

$$k_R(r) = \exp\left[-(e^\alpha r)^{e^\gamma} + \gamma + \alpha + (e^\gamma - 1)(\alpha + \log r)\right] \tag{26}$$

The numerical integration was conducted using stratified, antithetic sampling with strata defined by calendar year. Based on 100 repeated estimations, the standard deviation of the estimates was less than 7 percent of the average, estimated standard error. For all estimates, the first two decimals were stable. The maximum likelihood estimates obtained are presented below:

[Table 3 about here.]

The resulting gender specific estimates of the survivor function are shown in Figure 7, and the comparable non-parametric estimate is shown in Figure 8. The estimates show a rather skewed distribution with high probability mass for short durations and only little chance of durations of more than eighty years. The median duration is around 15 years and with males having shorter durations.

[Figure 7 about here.]

[Figure 8 about here.]

Combining this with the estimated incidence rate yields the projections of prevalence displayed in Figure 9 and 10. As always, prediction should only be done for the short term future, but as durations of diabetes are long it is reasonable to project prevalence twenty years into the future.

[Figure 9 about here.]

[Figure 10 about here.]

Regardless of estimation method, the projection is markedly larger than the observed prevalences on January 1 for the years 1993-2003 based on using a one year run-in period (Støvring, Andersen, Beck-Nielsen, Green, and Vach 2003). The prevalence is projected to decrease as a result of decreasing birth rates, cf. Figure 2, which is contrary to the markedly lower, but rapidly increasing prevalence observed throughout 1993-2003. The discrepancy between observed and projected prevalence reflects that past mortality among diabetics is severely underestimated and/or past incidence is severely overestimated, and so it clearly highlights the inadequacy of assuming stationarity. Yet the projection is still interesting from a public health perspective, since it is based on an assumption of stationary incidence and mortality. It thus describes the equilibrium size of the prevalent population if the current birth rate, incidence, and mortality among diseased are carried forward. In other words, if incidence and mortality remains at their current levels, then the number of prevalents will become more than doubled before a steady state is reached. This is in good agreement with other predictions of a doubling in prevalence over the next twenty years, both in the US (Mokdad, Bowman, Ford, Vinicor, Marks, and Koplan 2001) and worldwide (Wild, Roglic, Green, Sicree, and King 2004). If incidence further increases and/or mortality among diseased decreases, then even higher numbers of prevalent diabetics are to be expected. And it will take more than twenty years before equilibrium is reached.

5 DISCUSSION

In this paper we have developed and implemented methods for estimating and projecting incidence, prevalence and lifetime prevalence of a disease based on observation of incident events in an observation window, i.e. case cohort data. The developed methodology yields non-parametric estimates, and can likewise easily be applied if parametric estimates are desired.

In its simplest form, when assuming a stationary birth process and a stationary incidence, a simple non-parametric estimate of F is obtained. When alternatively the birth process is considered known, this is taken into account by a weighted, non-parametric estimate with weights based on the relative sizes of the relevant birth cohorts. Both approaches directly provide estimates of age-specific incidence as well as of lifetime risk, which are of considerable public health interest. Due to the relatively fast computational procedures developed, confidence intervals for the estimates could be obtained from direct application of bootstrap methodology.

The obtained projections demonstrate the lack of stationarity in the present situation of the diabetes epidemic, at least with respect to pharmacologically treated diabetes. With current incidence and mortality rates held constant, nearly a doubling in prevalence should be expected before a state of equilibrium is reached, which is predicted to not happen earlier than twenty years from now. If alternatively, the incidence rises and/or mortality among diseased goes down, then even higher prevalences are to be expected as well as a further postponement of equilibrium. Absence of stationarity in mortality is not surprising since insulin was introduced in the 1920's with virtually no prior treatment. Also, lack of stationarity of incidence would hardly come as a surprise with current focus on the impact of the changing lifestyle.

Although we in principle showed how the stationarity assumption could be relaxed by formulating a full, parametric likelihood, we did not give a detailed analysis of this situation due to its complexity. The data considered in this paper are rather limited since, first, the observation window is short compared to typical disase duration, and second, no information is available on age of onset outside the observation window. As a result, we have been unable to allow for trends in incidence and mortality, the absence of which must be considered most unrealistic—as also indicated by projections based on assumptions of stationarity. Still, we consider the results to be of interest, as they elicit the lack of equilibrium in current diabetes epidemiology, and outline how diabetes prevalence will continue to rise in the foreseeable future, even in the absence of rising incidence or declining mortality.

In many epidemiological settings it will, however, be possible to obtain data on age of onset for subjects prevalent at start of the time window or for diseased subjects dying in the observation window, cf. (Keiding, Holst, and Green 1989). Such information is obviously valuable and needs to be incorporated in the analysis to allow relaxing unrealistic assumptions. In such situations the data richness will, however, begin to resemble follow-up data, the lack of which motivated the present work.

References

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *J. R. Statist. Soc. A 154*(3), 371–412.

Keiding, N., C. Holst, and A. Green (1989, Sep). Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality. *Am J Epidemiol 130*(3), 588–600.

Lubin, J. and M. Gail (1984, Mar). Biased selection of controls for case-control analyses of cohort studies. *Biometrics 40*(1), 63–75.

Mantel, N. (1973, Sep). Synthetic retrospective studies and related topics. *Biometrics 29*(3), 479–86.

Mokdad, A., B. Bowman, E. Ford, F. Vinicor, J. Marks, and J. Koplan (2001). The continuing epidemics of obesity and diabetes in the United States. *JAMA 286*, 1195–200.

Oakes, D. (1981). Survival times: aspects of partial likelihood. *Internat. Statist. Rev. 49*(3), 235–264. With discussion and a reply by the author.

Prentice, R. L.; Breslow, N. E. (1978, April). Retrospective studies and failure time models. *Biometrika 65*(1), 153–158.

Prentice, R. L. (1986, April). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika 73*(1), 1–11.

Støvring, H., M. Andersen, H. Beck-Nielsen, A. Green, and W. Vach (2003). Rising prevalence of diabetes: evidence from a Danish pharmaco-epidemiological database. *The Lancet 362*, 537–38.

The WHO Collaborating Centre for Drug Statistics Methodology (2001). *ATC index with DDDs and Guidelines for ATC classification and DDD assignment*. Oslo.

Thomas, D. C. (1981, Dec). General relative-risk models for survival time and matched case-control analysis. *Biometrics 37*(4), 673–686.

Wild, S., G. Roglic, A. Green, R. Sicree, and H. King (2004, May). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care 27*(5), 1047–53.

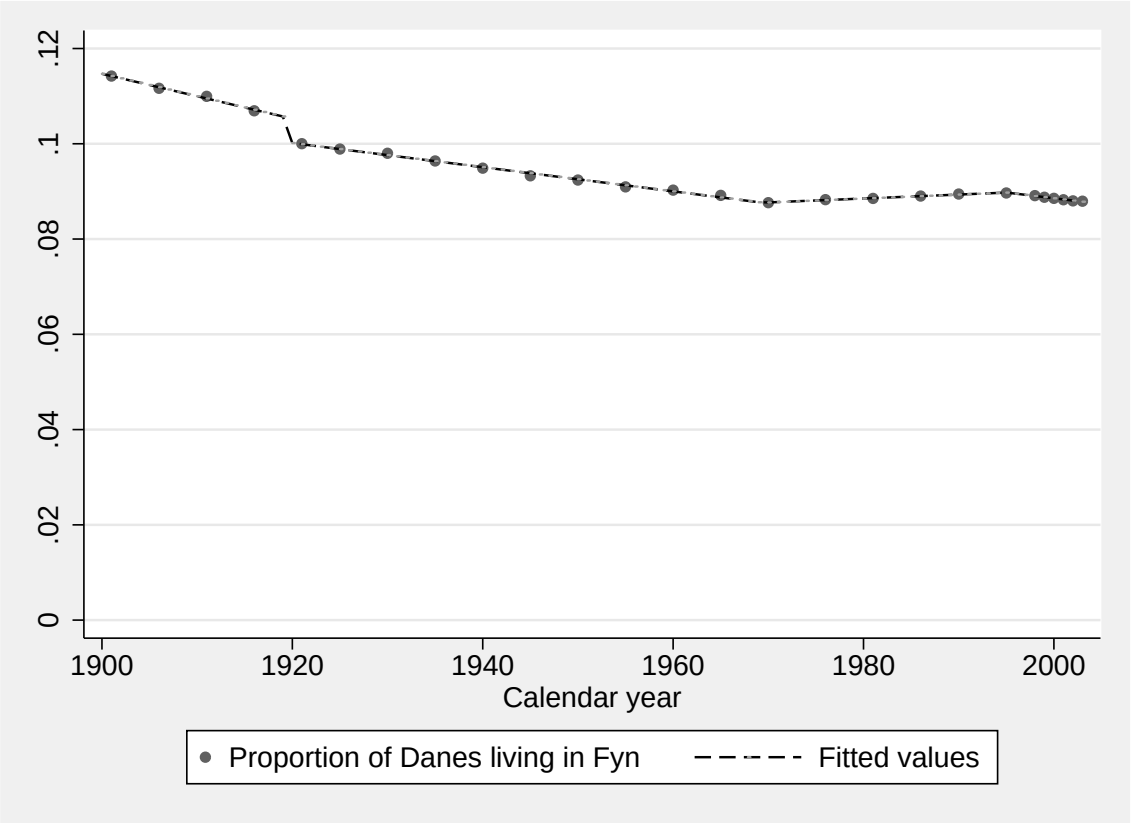

Figure 1: Observed and predicted fractions of the Danish population living in the county of Fyn during 1900-2003.

Figure 2: Annual number of births in the county of Fyn during 1891-2003.

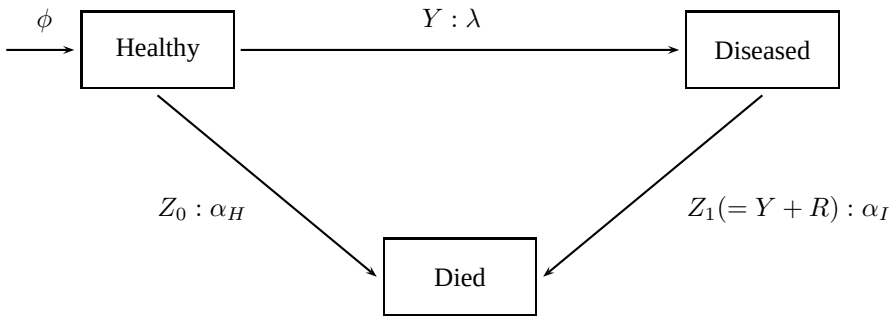

Figure 3: Illness-death model, where Y, Z_0 and Z_1 are random variables with associated hazards λ, α_H, and α_I, respectively.

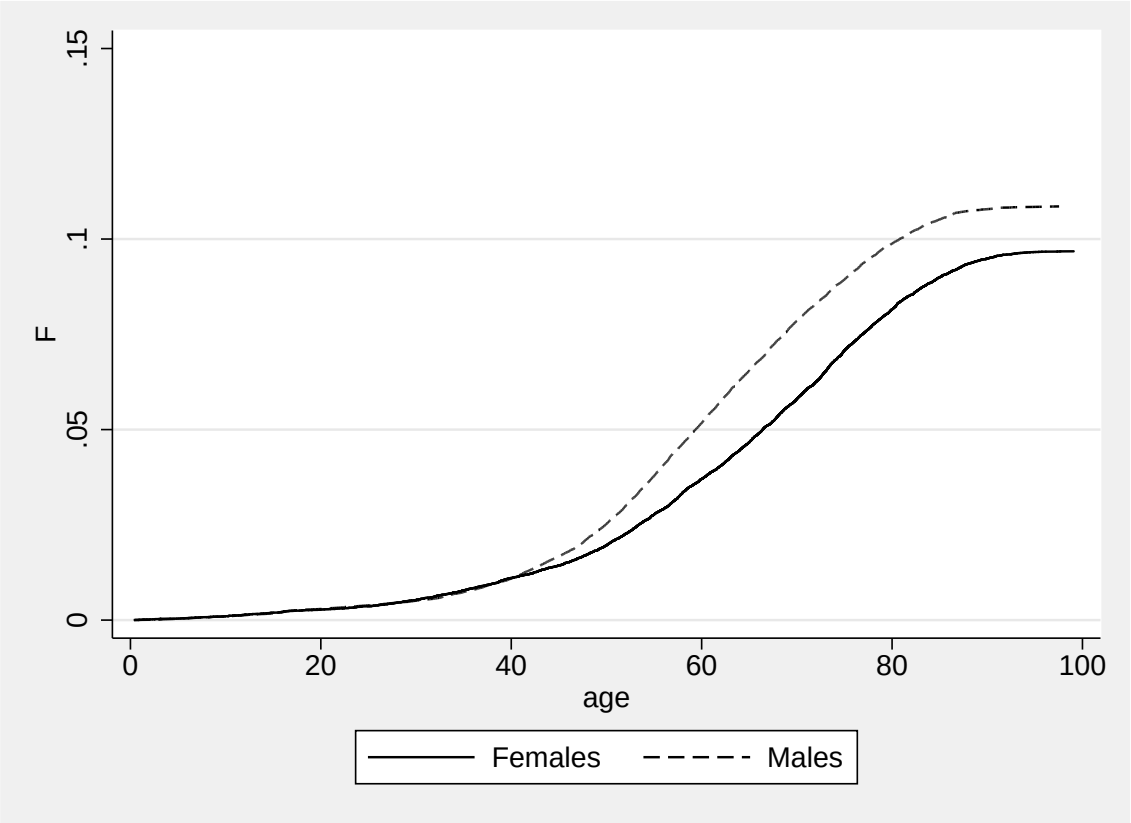

Figure 4: Estimated incidence distribution F for pharmacological treatment with any antidiabetic drug with respect to age and stratified on gender under the assumption of calendar time stationarity both with respect to incidence and birth process.

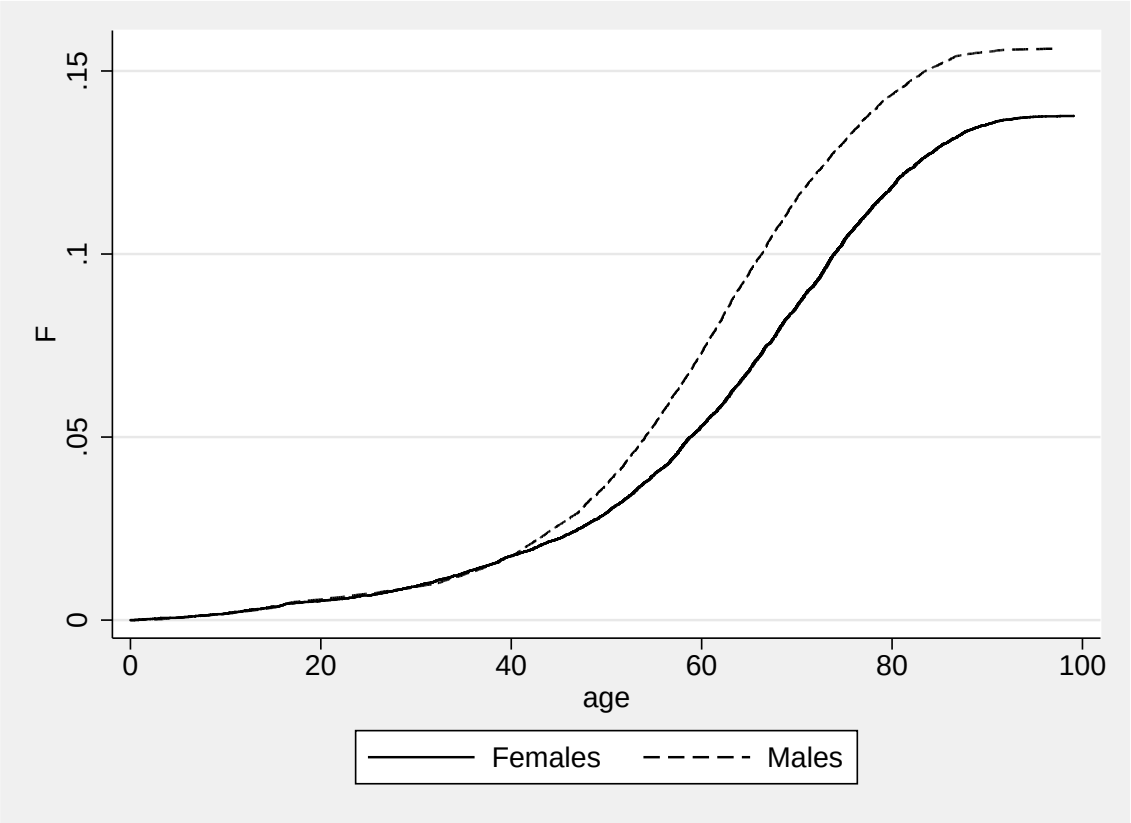

Figure 5: Estimated incidence distribution F for pharmacological treatment with any antidiabetic drug with respect to age and stratified on gender under the assumption of calendar time stationarity.

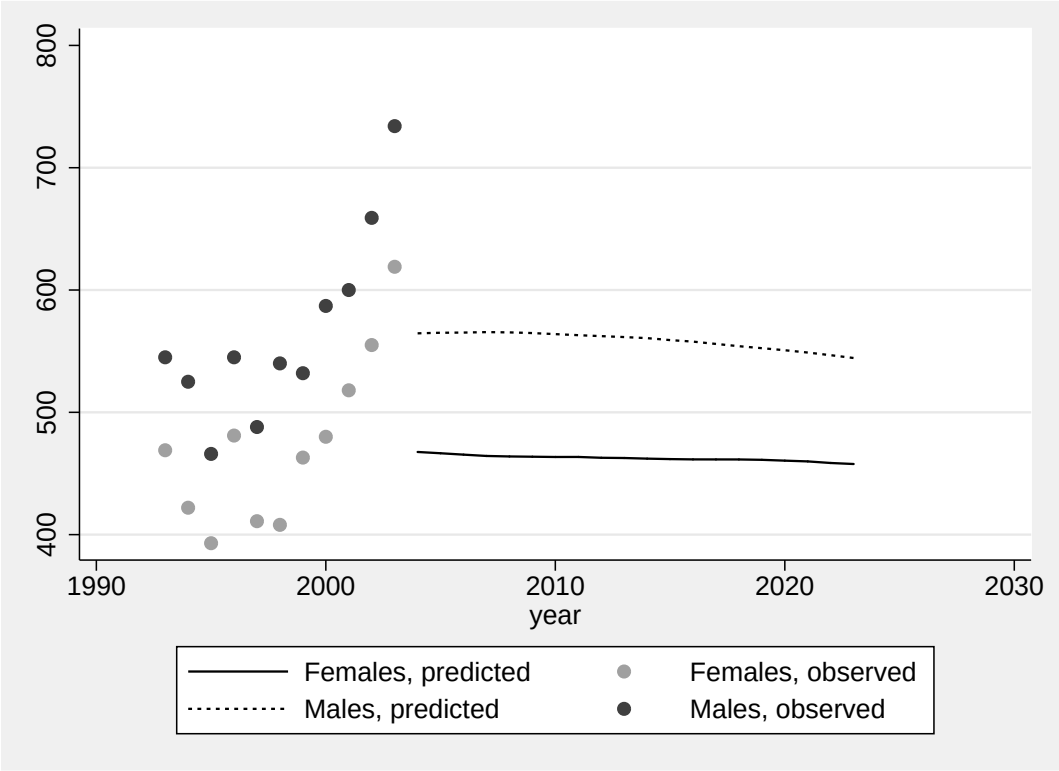

Figure 6: Projected and observed numbers of incident events based on an assumption of a stationary incidence and using a weighted, non-parametric estimate of F.

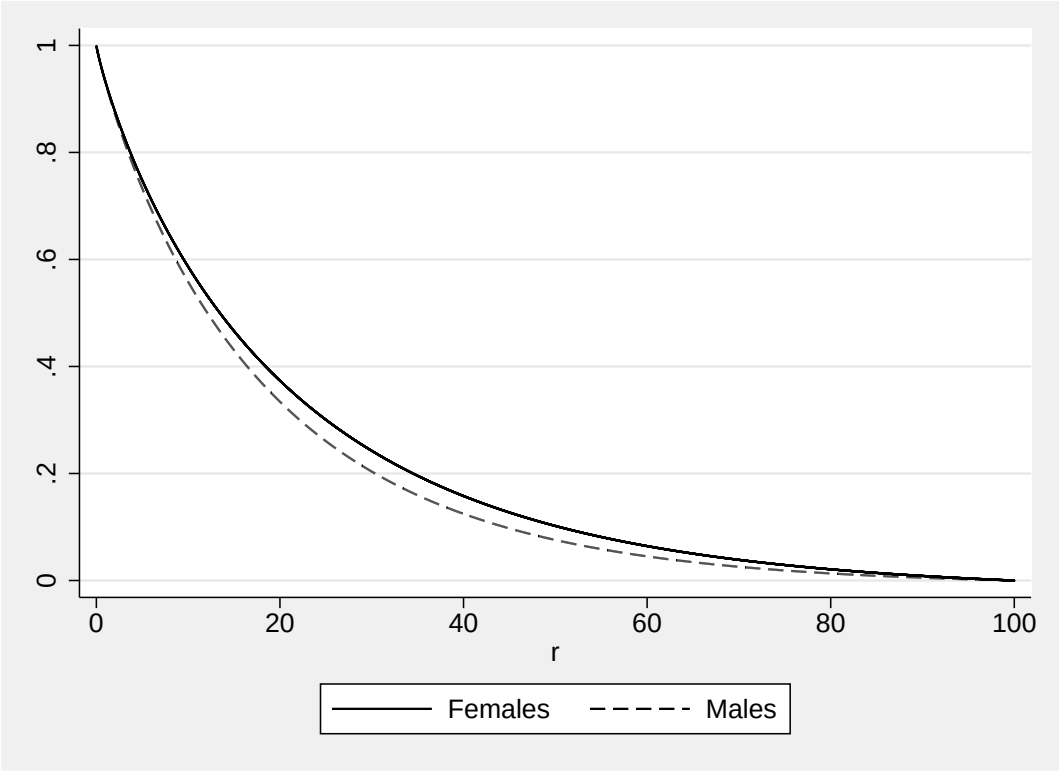

Figure 7: Parametric estimate of survivor function for duration of diabetes, Weibull.

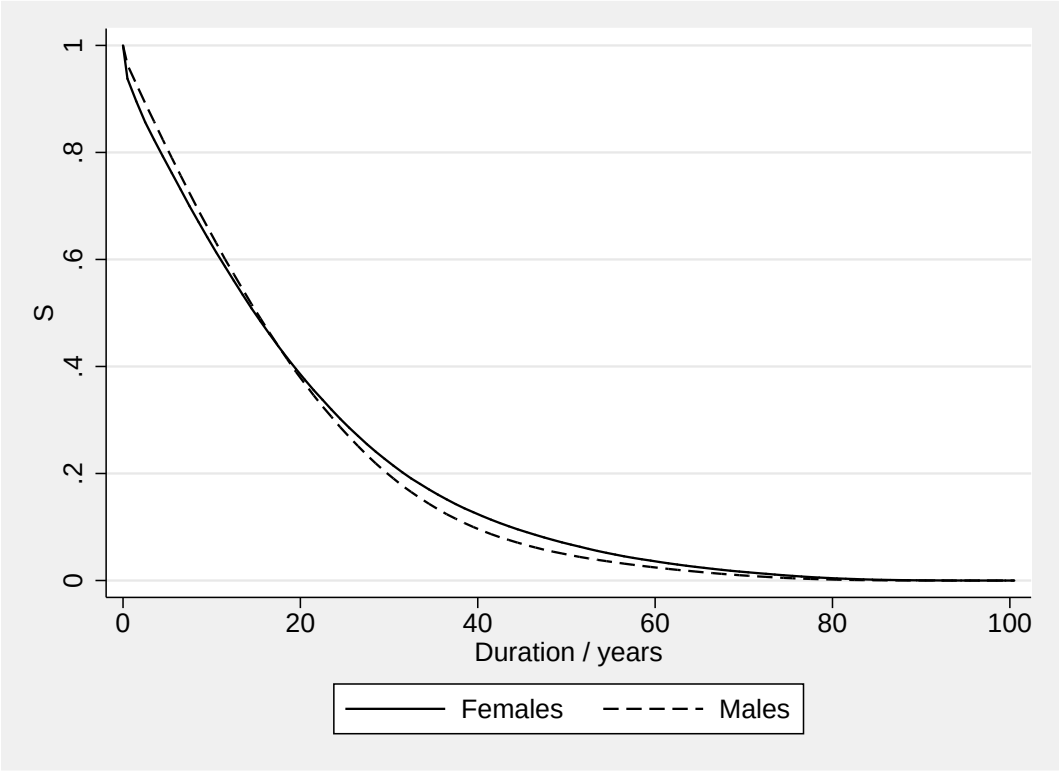

Figure 8: Non-parametric estimate of survivor function for duration of diabetes.

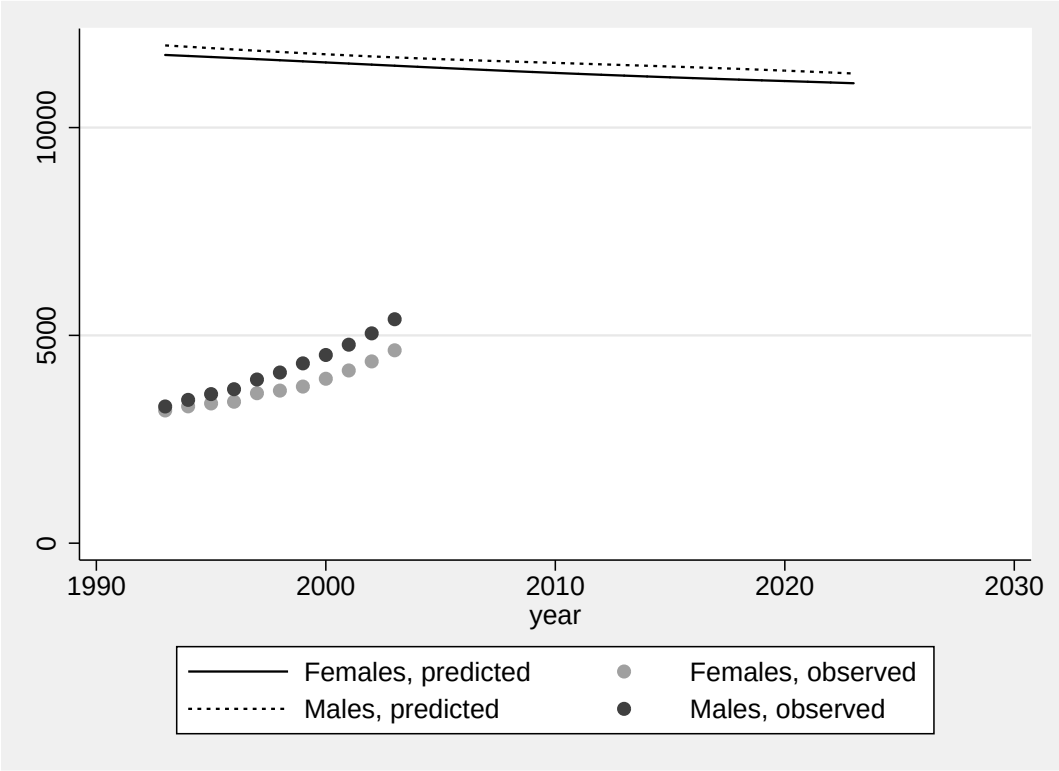

Figure 9: Projected and observed numbers of prevalent cases of diabetes, parametric estimate of K (Weibull).

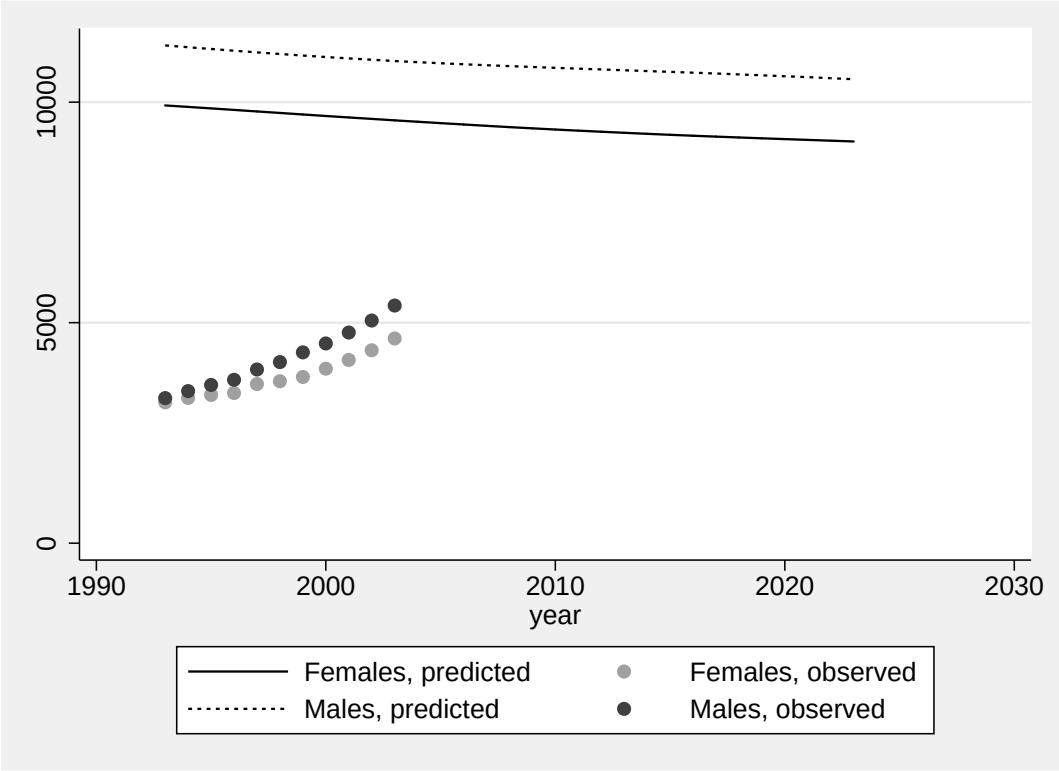

Figure 10: Projected and observed numbers of prevalent cases of diabetes, non-parametric estimate of K.

		Event year										
Gender	Birth	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Females	-1909	60	24	30	22	7	19	9	7	6	4	4
	1910-9	103	79	74	101	81	69	66	58	65	51	47
	1920-9	110	123	99	122	98	112	118	109	106	119	118
	1930-9	65	86	78	86	96	82	96	108	111	132	140
	1940-9	51	51	55	70	65	64	88	100	115	117	137
	1950-9	41	31	32	26	32	30	38	48	51	64	70
	1960-9	18	18	13	20	8	18	28	29	30	34	48
	1970-9	17	10	9	6	13	5	9	9	18	19	34
	1980-9	4		3	24	5	6	8	7	9	5	11
	1990-				4	6	3	3	5	7	10	10
Males	-1909	29	19	18	8	7	7	4		2	2	2
	1910-9	94	80	68	71	65	58	42	45	37	21	28
	1920-9	126	145	106	118	96	116	99	93	82	123	123
	1930-9	107	95	114	132	126	119	131	156	143	126	174
	1940-9	104	102	83	102	111	129	140	166	183	191	214
	1950-9	49	52	38	52	42	77	65	70	77	106	113
	1960-9	16	19	19	21	27	23	28	29	41	55	53
	1970-9	12	11	17	8	5	7	10	9	15	14	12
	1980-9	8	2	3	28	6	2	7	12	12	12	10
	1990-				5	3	2	6	7	8	9	5

Table 1: Number of incidence events by gender, calendar year of event and calendar year of birth.

	Females			Males		
Birth	r	$r+$	s	r	$r+$	s
-1909	163	29	378	91	7	188
1910-9	442	352	674	440	169	555
1920-9	351	883	433	493	734	514
1930-9	156	924	138	311	1,112	272
1940-9	61	852	39	147	1,378	120
1950-9	23	440	17	47	694	30
1960-9	5	259	7	7	324	15
1970-9	2	147	2	2	118	4
1980-9		82			102	1
1990-		48			45	

Table 2: Number of durations of diabetes treatment by gender, birth year, and observation type: observed (r), right censored ($r+$), and doubly truncated (s).

Gender	Parameter	Estimate	s.e.	95%-CI
Females	α	-3.054	0.048	(-3.148; -2.959)
	γ	-0.137	0.025	(-0.186; -0.087)
Males	α	-2.913	0.039	(-2.990; -2.837)
	γ	-0.101	0.022	(-0.144; -0.058)

Table 3: Gender specifi c parameter estimates for durations of diabetes, Weibull distribution.