Cancer Epidemiology 38 (2014) 193–199

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Determining disease prevalence from incidence and survival using simulation techniques

Simon Crouch*, Alex Smith, Dan Painter, Jinlei Li, Eve Roman

Epidemiology & Cancer Statistics Group, Department of Health Sciences, University of York, YO10 5DD, UK

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form 14 January 2014
Accepted 17 February 2014
Available online 18 March 2014

Keywords:
Prevalence
Incidence
Survival
Simulation
Monte-Carlo
Prevalence distribution

ABSTRACT

Objectives: We present a new method for determining prevalence estimates together with estimates of their precision, from incidence and survival data using Monte-Carlo simulation techniques. The algorithm also provides for the incidence process to be marked with the values of subject level covariates, facilitating calculation of the distribution of these variables in prevalent cases.

Methods: Disease incidence is modelled as a marked stochastic process and simulations are made from this process. For each simulated incident case, the probability of remaining in the prevalent sub-population is calculated from bootstrapped survival curves. This algorithm is used to determine the distribution of prevalence estimates and of the ancillary data associated with the marks of the incidence process. This is then used to determine prevalence estimates and estimates of the precision of these estimates, together with estimates of the distribution of ancillary variables in the prevalent sub-population. This technique is illustrated by determining the prevalence of acute myeloid leukaemia from data held in the Haematological Malignancy Research Network (HMRN). In addition, the precision of these estimates is determined and the age distribution of prevalent cases diagnosed within twenty years of the prevalence index date is calculated.

Conclusion: Determining prevalence estimates by using Monte-Carlo simulation techniques provides a means of calculation more flexible that traditional techniques. In addition to automatically providing precision estimates for the prevalence estimates, the distribution of any measured subject level variables can be calculated for the prevalent sub-population. Temporal changes in incidence and in survival offer no difficulties for the method.

1. Introduction

Estimation of disease prevalence is of fundamental interest in epidemiology [1]. As observed by Gigli et al. [2], three main methods of estimation of prevalence are commonly employed: cross-sectional population survey; direct count of cases in a disease register; and mathematical modelling based on incidence and survival rates. Gigli et al. [2] illustrate a method combining the latter two approaches in three steps: step 1 counts surviving cases at an index date from incident cases in a registry; step 2 estimates the number of prevalent cases lost-to-follow-up from the registry count; and step 3 estimates the number of prevalent cases at the index date that were incident before the start of the registry. Steps 1 and 2 together are often referred to as the "counting method" and step 3 as the "completeness index method".

Previous approaches to steps 2 and 3 of this schema have focussed on various analytic techniques of estimation [3–6], themselves based on the relationships between the various measurable quantities [7–9], or by direct modelling [10]. These techniques have found wide application in the literature [11–13]. Consideration of the precision of prevalence estimates has focussed on the variation implied by considering the incidence process as Poisson [2,14].

In this paper we will consider techniques of estimation for steps 2 and 3 based entirely on simulation. We will illustrate our techniques using data drawn from a population based cohort of patients diagnosed with haematological malignancies; in particular we will provide prevalence estimates for acute myeloid leukaemia (AML).

We first define what we mean by "prevalence". Broadly speaking the prevalence of a disease in a population is the number or proportion of the population alive at some index date,

* Corresponding author at: Epidemiology & Cancer Statistics Group, Department of Health Sciences, Seebohm-Rowntree Building, University of York, YO10 5DD, UK. Tel.: +44 01904 321938.
E-mail address: simon.crouch@ecsg.york.ac.uk (S. Crouch).

http://dx.doi.org/10.1016/j.canep.2014.02.005
1877-7821/© 2014 Elsevier Ltd. All rights reserved.

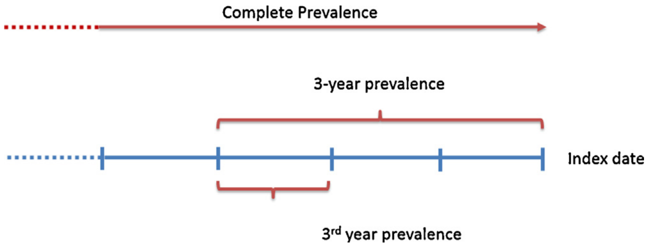

Fig. 1. Types of prevalence. Complete prevalence includes diagnosed cases from any time before the index date still in the prevalent population. *n*-year prevalence includes cases diagnosed within the last *n* years. *n*th year prevalence includes cases diagnosed during the single year between $(n-1)$ and *n* years before the index date.

previously diagnosed with the disease and not removed from the prevalent disease sub-population between diagnosis and the index date (by death or complete cure, for example); this is referred to as "complete prevalence". We also define "*n*-year prevalence" and "*n*th-year prevalence" to refer to those in the prevalent sub-population at the index date having received a diagnosis of the disease in the previous *n* years or during the *n*th year before the index date respectively. Therefore *n*-year prevalence is the sum of *k*th-year prevalence for *k* between 1 and *n*. So, for example, 3-year prevalence refers to all those in the prevalent sub-population on the index date diagnosed in the three years before the index date; 3rd-year prevalence refers to all those in the prevalent population on the index date diagnosed during the third year before the index date (Fig. 1). For simplicity of presentation in this paper, we assume that the only removal mechanism from the prevalent sub-population is death.

The relevance of *n*-year and *n*th-year prevalence for different values of *n* depends upon the use made of the estimates and upon the disease under consideration. *n*-year and *n*th-year prevalence estimates for small values of *n* will typically correspond to periods of intense treatment for acute diseases; for *n*th-year prevalence, larger values of *n* will correspond to periods of long term monitoring. For chronic diseases, all values of *n* are typically of interest.

In this paper we present a new method of determining prevalence based on the computationally intensive method of simulation. The advantages of this new method over existing methodology are that it naturally allows for the estimation of the precision of prevalence estimates and also allows for the estimation of ancillary information about the prevalent sub-population (for example, it allows for the estimation of the age distribution of the prevalent sub-population). In addition more complex modelling of incidence and survival functions than is usually allowed for in current techniques provides no additional obstacle to simulation techniques.

2. Materials

The determination by simulation of prevalence from incidence and survival estimates derived from a patient cohort is illustrated with data on patients diagnosed with acute myeloid leukaemia (AML) drawn from the UK's population-based Haematological Malignancy Research Network (HMRN) [15]. Initiated in 2004, and covering a population of 3.6 million, this unique patient cohort was established to provide robust generalizable data to inform clinical practice and research. Comprehensive information about HMRN is available elsewhere [15] but briefly, all patients newly diagnosed with a haematological malignancy residing in the HMRN region (>2000 patients a year) have full-treatment, response and outcome data collected to clinical trial standards. HMRN has Section 251 support under the NHS Act 2006; enabling the Health and Social Care Information Centre (HSCIC) to routinely link to and release nationwide information on deaths, subsequent cancer registrations, and Hospital Episode Statistics (HES). Loss-to-follow-up rates are very low in this registry, thanks to this comprehensive data linkage. In fact, for the small number of subjects that are lost-to-follow-up (by emigration from the UK, for example), the actual date of loss is known with precision in this registry. The demographic structure of the region is similar to the demographic structure of the UK as a whole, allowing for reliable generalization from this population to the population of the UK.

Incidence data on patients, 18 years and older, diagnosed with AML was available from the HMRN registry for seven years from 01/09/2005 to 31/08/2012. The index date for the calculations of prevalence was taken to be 31/08/2011 and years are taken to run from the first of September to the thirty-first of August. Survival outcome data was available up until 26/03/2013 for patients diagnosed between 01/09/2005 and 31/08/2011. Characteristics of these patients are shown in Table 1.

2.1. Methods

We can estimate the number of prevalent cases of a disease at a particular index date by combining information on incidence and survival. An incident case at time *t* before the index date, characterized by a vector of explanatory variables for survival $\mathbf{x}$, will contribute $S(\mathbf{x}, t)$ to the expected number of prevalent cases at the index date, where $S(\mathbf{x}, t)$ is the survival function conditional on explanatory variables $\mathbf{x}$. Therefore, if there are *n* cases incident at times $\{t_1, t_2, \ldots, t_n\}$ each with corresponding survival explanatory variables $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, then the expected number of prevalent cases at the index date T_0 is given by

$$P = \sum_{i=1}^{n} S(\mathbf{x}_i, T_0 - t_i)$$

In this paper we take prevalent cases to be those that have ever been diagnosed with the disease under consideration. This can easily be generalized so that prevalence refers to subjects that have not been removed from the prevalent sub-population by other means (such as cure) by taking the end-point for the survival function *S* to be time to removal from prevalent population rather than simply time to death. Complete prevalence takes the sum over all time before the index date; this generalizes to estimation of *n*-year and *n*th-year prevalence by restricting the sum to cover incident cases from the corresponding time period.

The value of *P* can be calculated by simulation. If the times and associated survival explanatory variables of incident cases can be appropriately modelled, and if survival conditional on the explanatory variables can be estimated, then simulation from the incidence model, together with the survival function, will provide an estimate of *P*. What is more, sources of variation can be taken into account, so that calculations of *P* from repeated random draws from the incidence model will provide an estimate of the

Table 1
Characteristics of the AML patient cohort.

	Incidence dataset	Survival dataset
Incidence dates	01/09/2005–31/08/2012	01/09/2005–31/08/2011
Number of subjects	1079	934
Male	592 (55%)	517 (55%)
Female	487 (45%)	417 (45%)
Age range (years)	18.7–97.8	19.0–97.8
Median age (IQR)	71.9 (60.0–79.8)	71.7 (60.0–79.3)
Maximum follow-up (days)	N/A	2720
Median survival (95% CI)	N/A	132 (109–159)
Total follow-up (years)	N/A	1390

IQR, inter quartile range.