Anna Gigli

THE VARIANCE OF THE COMPLETENESS INDEX

W.P. 3/2001
dicembre 2001

Istituto di Ricerche sulla Popolazione, IRP-CNR
via Nizza 128, 00198 Roma

Abstract

Cancer prevalence is the proportion of people in a population diagnosed with cancer in the past and still alive. One way to estimate prevalence is via population-based registries, where data on diagnosis and life status of all incident cases occurring in the covered population are collected, as described in Gigli *et al.* (2000).

This paper describes in more details the analytical approach to the estimation of the variance of the completeness index, which in turn allows to compute the variance of the complete prevalence. [1]

keywords: limited duration prevalence, complete prevalence, completeness index, variance estimation, cancer registries

1 Introduction

Complete prevalence of a chronic disease is the proportion of people alive on a certain date who previously had been diagnosed of the disease, regardless of how long ago it was diagnosed.

For populations covered by disease registration, data on diagnosis and life status of all incident cases collected by the registry provide the most reliable basis for the calculation of prevalence. In a population-based registry all disease incidence cases occurring in the covered population are registered.

The most direct method to estimate prevalence from registry data is the counting method, which consists in counting the number of registered cases that survive at any specified time and age. In practice, however, the estimation is complicated by several phenomena such as migrations, cases registered only at death, cases lost to follow-up, cases diagnosed before the start of registration.

Correction for the first three phenomena yields to the so-called *limited duration prevalence*. The last one, however, is the most important for newly-estabilished registries, and causes a bias in the prevalence estimation whose extent depends on the length of the registration period and the shape of the patient incidence and survival functions. Capocaccia and De Angelis (1997) developed a method to account for this bias, which consists of estimating a

[1] This manuscript is part of a research contract with the National Cancer Institute, Bethesda, USA, order no. 263-MQ-117231-1 of September 25, 2001.

correction factor named *completeness index*. By dividing the limited duration prevalence by the completeness index we obtain what we call the *complete prevalence*.

In order to compute the variance of the complete prevalence we need to estimate the variance of the completeness index, and this is the aim of this work.

The report is structured as follows. Section 2 introduces definitions and notations; section 3 illustrates the general method to compute the completeness index variance; sections 4 and 5 describe the survival and incidence models used in the estimation of the completeness index; section 6 explains the algorithm for the computation of the completeness index variance; finally section 7 suggests some further developments.

2 Definitions

In a birth cohort c let us define the prevalence $N(x)$ as the proportion in a population of individuals of age x alive on a certain date who previously had been diagnosed of the disease; when the prevalence is computed regardless of how long ago the disease was diagnosed, it is called *complete prevalence*.

When the disease is irreversible and assuming that the birth process is independent of the subsequent life stories, there is a simple relationship between prevalence, incidence and survival, given by the convolution

$$N(x) = \int_0^x I(t)S(x-t,t)dt$$

where $I(t)$ is the incidence hazard at age t and $S(x-t,t)$ is the probability that individuals diagnosed with cancer at age t are still alive at age x.

The estimation of N is particularly simple if I and S are known parametric functions: if ψ denotes the parameter vector the resulting *modelled prevalence* will be denoted by $N(x;\psi)$.

If L is the lenght time (in years) the registry has been operating, the modelled prevalence $N(x;\psi)$ decomposes into two parts: $N_O(x,L;\psi)$, the (modelled) observed prevalence at age x of incident cases registered in age interval $[x-L,x]$ and $N_U(x,L;\psi)$, the (modelled) unobserved prevalence at age x of unregistered cases diagnosed in age interval $[0,x-L]$ and still alive

$$N(x;\psi) = N_O(x,L;\psi) + N_U(x,L;\psi),$$

where

$$N_O(x, L; \psi) = \int_{x-L}^{x} I(t; \psi) S(x - t, t; \psi) dt$$

and

$$N_U(x, L; \psi) = \int_{0}^{x-L} I(t; \psi) S(x - t, t; \psi) dt.$$

The *completeness index* R is the proportion of modelled prevalence which is observed, and is defined as

$$R(x, L; \psi) = \frac{N_O(x, L; \psi)}{N(x; \psi)}. \tag{1}$$

3 The computation of $\mathrm{var}(R)$

Let $\hat{\psi}$ be the vector of the maximum likelihood estimates of the incidence and survival parameter vector ψ and let $\hat{V}$ be the estimated covariance matrix of $\hat{\psi}$. An approximation to the variance of R can be calculated by applying the delta method:

$$\mathrm{var}[R(x, L; \hat{\psi})] \approx \left(\frac{\partial R}{\partial \psi} \Big|_{\psi=\hat{\psi}} \right)^T \hat{V} \left(\frac{\partial R}{\partial \psi} \Big|_{\psi=\hat{\psi}} \right) \tag{2}$$

where $\partial R / \partial \psi$ is the vector of partial derivatives of R with respect to the incidence and survival parameters.

The first step for solving (2) is the computation of $\partial R / \partial \psi$, for each component of ψ

$$\frac{\partial R}{\partial \psi_i} = \frac{\frac{\partial}{\partial \psi_i} N_O(x, L; \psi) - R(x, L; \psi) \frac{\partial}{\partial \psi_i} N(x; \psi)}{N(x; \psi)}. \tag{3}$$

We need to calculate the two derivatives with respect to each component of the vector of the parameter estimates. The derivatives of $N(x; \psi)$ are

$$\frac{\partial}{\partial \psi_i} N(x; \psi) = \int_0^x \frac{\partial}{\partial \psi_i} \left[I(t; \psi) S(x - t, t; \psi) \right] dt. \tag{4}$$

Similarly for N_O, where only the integration limits change.

Once the integrand in (4) is computed, we estimate the correspondent integral via the Simpson method and obtain $\frac{\partial}{\partial \psi_i} N(x; \psi)$ and $\frac{\partial}{\partial \psi_i} N_O(x; \psi)$.

The results are then plugged into (3) and $\frac{\partial R}{\partial \psi_i}$ are obtained for each parameter. We substitute (3) into (2) and obtain the variance of the completeness index.

4 Survival model

Let

$$S(x-t,t;\psi) = \{(1-Q) + Q\exp[-[\lambda(x-t)]^{\beta}]\}^{\exp[\gamma_1(t-t_0)+\gamma_2(t+c-s_0)+\gamma_3\delta]}. \quad (5)$$

be the model for the cumulative relative survival function for a patient diagnosed in year $t+c$ at age t, who survives until age x (De Angelis *et al.* , 1999). This class of models assumes that only a portion Q of the individuals with cancer will die with a relative survival following a Weibull distribution with parameters λ, β, while the remaining $(1-Q)$ have the same mortality rate as the general population, and consequently their relative survival is 1. In the model considered here the death hazard due to cancer is assumed to be linearly dependent, on a logarithmic scale, of age at diagnosis and calendar time of diagnosis; the corrections due to the risk of being diagnosed one year older than the reference age t_0 and one year later than the reference year s_0 are parameterized by γ_1 and γ_2, respectively (we take t_0 as the mean age at diagnosis, which varies according to the different cancer site, and s_0 as the year 1983.5); finally the reference race/ethnicity is white and the correction due to the risk of belonging to black or hispanic race/ethnicity is parameterized as γ_3, while δ is a dummy variable for the race/ethnicity.

5 Incidence models

For the incidence hazard we distinguish two models: the exponential model, based on the strong assumption of independence between age and cohort, which has been shown to have a biological rationale for a general class of cancers (Armitage and Doll, 1954); and the polynomial model which applies to cancers whose growth depends on hormonal factors (Capocaccia *et al.* , 1990).

5.1 Exponential incidence

Let

$$\mathrm{I}(x;\psi) = \exp(a_c)x^b \tag{6}$$

be the cancer incidence for a person of current age x who belongs to the $c-th$ birth cohort. In presence of rare events (such as cancer) $I \to 0$ and (6) can be approximated by

$$\mathrm{I}(x;\psi) = \frac{1}{1+\exp\{-[a_c + b\log(x)]\}},$$

which in turn leads to

$$\mathrm{logit}(\mathrm{I}(x;\psi)) = a_c + b\log(x).$$

The parameter b is to be estimated, while the variable a_c, which depends on the birth cohort, is a multiplicative variable which will cancel out in the ratio (1) and therefore in what follows will not be considered.

5.2 Polynomial incidence

Various studies have reported that, in those cancers for which the polynomial model applies, the best fit of the incidence hazard is obtained by a 6-degree polynomial. Here we use the notation proposed by Merrill *et al.* (2000): let

$$I(x;\psi) = \frac{1}{1+\exp\{-[a_c + b_1(x-t_1) + b_2(x-t_1)^2 + \ldots + b_6(x-t_1)^6]\}}, \tag{7}$$

where a_c is the logit of the incidence at the c-th birth cohort, when age is equal to the reference age t_1, $b_1, \ldots, b_6$ are the parameters to be estimated, together with a_c, which for polynomial incidence does not cancel out in the ratio (1). The corresponding logit is

$$\mathrm{logit}(I(x;\psi) = a_c + b_1(x-t_1) + b_2(x-t_1)^2 + \ldots + b_6(x-t_1)^6.$$

6 Partial derivatives of I $*$ S

6.1 Exponential incidence

From models (5) and (6) we obtain the parameter vector $\psi = (Q, \lambda, \beta, \gamma_1, \gamma_2, \gamma_3, b)$, where the incidence parameter b is independent of the survival parameters.

Let

$$F(t,x;\psi) = I(t;\psi) * S(x-t,t;\psi) = t^b\{1-Q+Qe^{-\kappa_1}\}^{\exp(\kappa_2)},$$

where
$\kappa_1 = [\lambda(x-t)]^\beta$,
$\kappa_2 = \gamma_1(t-t_0) + \gamma_2(t+c-s_0) + \gamma_3\delta$,
and let $F_i(t,x;\psi) = \frac{\partial}{\partial\psi_i}F(t,x;\psi)$, for $i = 1,\ldots,7$.
We have

$$\begin{aligned}
F_1(t,x;\psi) &= \frac{\partial}{\partial Q}F(t,x;\psi) = F(t,x;\psi) * \frac{e^{\kappa_2}(e^{-\kappa_1}-1)}{1-Q+Qe^{-\kappa_1}} \\
F_2(t,x,\psi) &= \frac{\partial}{\partial\lambda}F(t,x;\psi) = F(t,x,\psi) * \frac{-Q\kappa_1\beta e^{\kappa_2-\kappa_1}}{(1-Q+Qe^{-\kappa_1})\lambda} \\
F_3(t,x,\psi) &= \frac{\partial}{\partial\beta}F(t,x;\psi) = F(t,x;\psi) * \frac{-Q\kappa_1\log(\kappa_1)e^{\kappa_2-\kappa_1}}{(1-Q+Qe^{-\kappa_1})\beta} \\
F_4(t,x,\psi) &= \frac{\partial}{\partial\gamma_1}F(t,x;\psi) = F(t,x;\psi) * e^{\kappa_2}(t-t_0)\log(1-Q+Qe^{-\kappa_1}) \\
F_5(t,x,\psi) &= \frac{\partial}{\partial\gamma_2}F(t,x;\psi) = F(t,x;\psi) * e^{\kappa_2}(t+c-s_0)\log(1-Q+Qe^{-\kappa_1}) \\
F_6(t,x,\psi) &= \frac{\partial}{\partial\gamma_3}F(t,x;\psi) = F(t,x;\psi) * e^{\kappa_2}\delta\log(1-Q+Qe^{-\kappa_1}) \\
F_7(t,x;\psi) &= \frac{\partial}{\partial b}F(t,x;\psi) = F(t,x;\psi) * \log(t).
\end{aligned}$$

6.2 Polynomial incidence

From models (5) and (7) we obtain 6 survival and 7 incidence parameters, mutually orthogonal to each other, and the parameter vector is $\psi = (Q,\lambda,\beta,\gamma_1,\gamma_2,\gamma_3,a_c,b_1,\ldots,b_6)$.
Let

$$F(t,x;\psi) = I(t;\psi)S(x-t,t;\psi) = \frac{\{1-Q+Qe^{-\kappa_1}\}^{\exp(\kappa_2)}}{1+\exp(-\kappa_3)},$$

where
$\kappa_1 = [\lambda(x-t)]^\beta$,
$\kappa_2 = \gamma_1(t-t_0) + \gamma_2(t+c-s_0) + \gamma_3\delta$,
$\kappa_3 = a_c + b_1(x-t_1) + b_2(x-t_1)^2 + \ldots + b_6(x-t_1)^6$,

and let $F_i(t,x;\psi) = \frac{\partial}{\partial \psi_i} F(t,x;\psi)$, for $i = 1, \ldots, 13$.
We have

$$
\begin{aligned}
F_1(t,x;\psi) &= \frac{\partial}{\partial Q} F(t,x;\psi) = F(t,x;\psi) * \frac{e^{\kappa_2}(e^{-\kappa_1}-1)}{1-Q+Qe^{-\kappa_1}} \\
F_2(t,x,\psi) &= \frac{\partial}{\partial \lambda} F(t,x;\psi) = F(t,x,\psi) * \frac{-Q\kappa_1\beta e^{\kappa_2-\kappa_1}}{(1-Q+Qe^{-\kappa_1})\lambda} \\
F_3(t,x,\psi) &= \frac{\partial}{\partial \beta} F(t,x;\psi) = F(t,x;\psi) * \frac{-Q\kappa_1 \log(\kappa_1) e^{\kappa_2-\kappa_1}}{(1-Q+Qe^{-\kappa_1})\beta} \\
F_4(t,x,\psi) &= \frac{\partial}{\partial \gamma_1} F(t,x;\psi) = F(t,x;\psi) * e^{\kappa_2}(t-t_0)\log(1-Q+Qe^{-\kappa_1}) \\
F_5(t,x,\psi) &= \frac{\partial}{\partial \gamma_2} F(t,x;\psi) = F(t,x;\psi) * e^{\kappa_2}(t+c-s_0)\log(1-Q+Qe^{-\kappa_1}) \\
F_6(t,x,\psi) &= \frac{\partial}{\partial \gamma_3} F(t,x;\psi) = F(t,x;\psi) * e^{\kappa_2}\delta\log(1-Q+Qe^{-\kappa_1}) \\
F_7(t,x;\psi) &= \frac{\partial}{\partial a_c} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * \exp(-\kappa_3) \\
F_8(t,x;\psi) &= \frac{\partial}{\partial b_1} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * (x-t_1)\exp(-\kappa_3) \\
F_9(t,x;\psi) &= \frac{\partial}{\partial b_2} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * (x-t_1)^2\exp(-\kappa_3) \\
F_{10}(t,x;\psi) &= \frac{\partial}{\partial b_3} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * (x-t_1)^3\exp(-\kappa_3) \\
F_{11}(t,x;\psi) &= \frac{\partial}{\partial b_4} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * (x-t_1)^4\exp(-\kappa_3) \\
F_{12}(t,x;\psi) &= \frac{\partial}{\partial b_5} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * (x-t_1)^5\exp(-\kappa_3) \\
F_{13}(t,x;\psi) &= \frac{\partial}{\partial b_6} F(t,x;\psi) = F(t,x;\psi) * I(t;\psi) * (x-t_1)^6\exp(-\kappa_3)
\end{aligned}
\tag{8}
$$

7 Further developments

In those cases where the polynomial model fits the incidence hazard better than the exponential model, the 6-degree polynomial could be approximated by a cubic spline, which provides smoother estimated curves. In particular

the use of a restricted cubic spline, which is forced to be linear before the first and after the last knot, seems suitable to deal with the problem of data scarcity at both ends of the age interval.

The algorithm for the computation of the derivatives of $I * S$ does not change in its structure, but a special care should be paid in the choice of the knots.

Acknowledgement

This manuscript has been prepared when the author was still working at the Istituto per le Applicazioni del Calcolo, IAC-CNR; it is part of a research contract with the National Cancer Institute, Bethesda, USA, order no. 263-MQ-117231-1 of September 25, 2001.

References

P. Armitage and R. Doll (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, 1–15.

R. Capocaccia and R. De Angelis (1997). Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine*, **16**, 425–440.

R. Capocaccia, A. Verdecchia, A. Micheli, M. Saint, G. Gatta, F. Berrino (1990). Breast cancer incidence and prevalence estimated from survival and mortality. *Cancer Causes Control*, **1**, 23–29.

R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, A. Verdecchia (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, **18**, 441–454.

A. Gigli, A. Mariotto, I. Corazziari, R. Capocaccia (2000). Estimating the variance of chronic disease prevalence. *Quaderno IAC*, **24/2000.**

R.M. Merrill, R. Capocaccia, E.J. Feuer, A. Mariotto (2000) 'Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program', *International Journal of Epidemiology*, **29**, 197–207.