Istituto di Ricerche sulla Popolazione
e le Politiche Sociali - CNR

IRPPS Working Papers

ISSN 2240-7332

Variance of complete cancer prevalence at all ages

Anna Gigli

What is IRPPS?

IRPPS is an Interdisciplinary Research Institute that conducts studies on demographic and migration issues, welfare systems and social policies, on policies regarding science, technology and higher education, on the relations between science and society, as well as on the creation of, access to and dissemination of knowledge and information technology.

www.irpps.cnr.it

IRPPS WPs n. 14 (2007)

Variance of complete cancer prevalence at all ages

Anna Gigli

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.

In this paper a method for the estimation of the variance of complete prevalence at all ages combined has been investigated. The proposed solution consists of estimating an upper bound of the variance of interest. Simulations show that the upper bound works well, however a theoretical proof has not been found.

The paper is organized as follows: after a brief introduction in section 1, the problem is illustrated in section 2; a new solution is proposed in section 3 and applied to a simulated data set in section 4; and finally some technical advice on how to modify the existing software is proposed in section 5.

Keywords: Complete prevalence, Variance estimation, Cancer registries, Incidence, survival, SEER*Stat software

Citazione consigliata:

Gigli, Anna. Variance of complete cancer prevalence at all ages. *IRPPS Working Papers*, n. 14, 2007.

Anna Gigli è ricercatrice e presso presso l'Istituto Superiore di Sanità (e-mail: anna.gigli@irpps.cnr.it).

Istituto di Ricerche sulla Popolazione e le Politiche Sociali - CNR
Via Palestro, 32 - 00185 Roma
http://www.irpps.cnr.it/it

1 Introduction

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute reports annually the number of persons alive following a diagnosis of cancer, or complete prevalence of cancer. This statistic is estimated in 2 steps:

1. The number of alive cancer cases reported to the SEER data after 1975 are counted and survivors among people lost to follow-up are estimated and added. This is denoted limited duration prevalence and can be calculated using the SEER*Stat software.

2. The proportion of prevalence that is unobserved, i.e. prevalence of cases diagnosed prior to 1975, is estimated using the completeness index method. Limited duration prevalence is adjusted to represent complete prevalence or lifetime prevalence. The completeness index method (Capocaccia and De Angelis, 1997) is implemented into COMPREV, a new software that calculates complete prevalence by adjusting limited duration prevalence imported from SEER*Stat with the completeness index.

Variance of limited duration prevalence is given in Clegg *et al.*(2002), variance of the completeness index and consequently of complete prevalence is given in Gigli *et al.*(2006). In both cases the variance is provided in 5-year age groups.

Extending the variance of complete prevalence to all age-groups requires an extra effort, since it implies estimating some covariance matrices. This is the aim of the present work.

2 The problem

We briefly recall some notation; further details can be found in Gigli *et al.*(2006).

For a fixed birth cohort c and a fixed age at prevalence x

- *complete prevalence* $N_x(0,x)$ is defined as the portion of people aged x alive on a certain date who had been diagnosed of the disease between ages 0 and x;

- *limited duration prevalence* $\tilde{N}_x(x-L,x)$ is the prevalence at age x estimated by population-based cancer registries and is based on a limited observational period L (Gail *et al.*, 1999);

- *modelled prevalence* $\hat{N}_x(0,x;\underline{\hat{\psi}})$ is a parametric estimate of the prevalence at age x, based on a complex function (convolution) of incidence and survival parametric models; $\underline{\hat{\psi}}$ is the maximum likelihood estimate of the vector of the incidence and survival parameters (Verdecchia *et al.*, 1989);

- *completeness index* $R_x(L;\underline{\hat{\psi}}) = \dfrac{\hat{N}_x(x-L,x;\underline{\hat{\psi}})}{\hat{N}_x(0,x;\underline{\hat{\psi}})}$ is the proportion of modelled prevalence at age x that is observed (Capocaccia and De Angelis, 1997);

- an estimate of complete prevalence at age x is obtained by combining limited duration prevalence and completeness index

$$N_x^*(0,x;\underline{\hat{\psi}}) = \frac{\tilde{N}_x(x-L,x)}{R_x(L;\underline{\hat{\psi}})} \tag{1}$$

- the analytical approximation to the variance of the completeness index is

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

  where $\hat{\boldsymbol{V}}$ is the covariance matrix of the mle vector $\hat{\psi}$ and $\partial R_x/\partial\underline{\psi}$ is the vector of partial derivatives of R_x with respect to the components of the parameter vector $\underline{\psi}$ (Gigli *et al.*, 2006).

We want to compute the variance of the estimated complete prevalence for all ages

$$\begin{aligned}\mathrm{var}(N^*) &= \mathrm{var}\left(\sum_{x=0}^{age_{max}} \frac{\tilde{N}_x}{R_x}\right) \\ &= \sum_{x=0}^{age_{max}} \mathrm{var}\left(\frac{\tilde{N}_x}{R_x}\right) + \sum_{x\neq y;x,y=0}^{age_{max}} \mathrm{cov}\left(\frac{\tilde{N}_x}{R_x}, \frac{\tilde{N}_y}{R_y}\right),\end{aligned} \tag{3}$$

knowing that:

$$\mathrm{cov}(R_x, R_y) \approx \left(\frac{\partial R_x}{\partial \underline{\psi}}\Big|_{\underline{\psi}=\underline{\hat{\psi}}}\right)^T \hat{\boldsymbol{V}} \left(\frac{\partial R_y}{\partial \underline{\psi}}\Big|_{\underline{\psi}=\underline{\hat{\psi}}}\right) \tag{4}$$

(an extension of (2)), and

$$\mathrm{var}\left(\frac{\tilde{N}_x}{R_x}\right) \approx \frac{1}{R_x^2}\mathrm{var}(\tilde{N}_x) - 2\frac{\tilde{N}_x}{R_x^3}\mathrm{cov}(\tilde{N}_x, R_x) + \frac{\tilde{N}_x^2}{R_x^4}\mathrm{var}(R_x) \tag{5}$$

(Gigli *et al.*, 2006).

3 A solution

Let

- $\underline{\omega} = (\tilde{N}_x, R_x, \tilde{N}_y, R_y)$ be a vector of 4 components and let $\underline{\hat{\omega}}$ be the vector of estimates;
- $g(\underline{\omega}) = \frac{\tilde{N}_x}{R_x}$ and $h(\underline{\omega}) = \frac{\tilde{N}_y}{R_y}$ be two different functions of the vector $\underline{\omega}$. Notice that the random variables $\tilde{N}_x \neq \tilde{N}_y$ and $R_x \neq R_y$, hence the two functions g and h are distinct.

We compute $\mathrm{cov}(g(\underline{\omega}), h(\underline{\omega}))$ in terms of $\mathrm{cov}(\underline{\omega})$, as in (4):

$$\mathrm{cov}(g(\underline{\omega}), h(\underline{\omega})) \approx \left(\frac{\partial g}{\partial \underline{\omega}}\Big|_{\underline{\omega}=\underline{\hat{\omega}}}\right)^T \mathrm{cov}(\underline{\omega}) \left(\frac{\partial h}{\partial \underline{\omega}}\Big|_{\underline{\omega}=\underline{\hat{\omega}}}\right),$$

where $\frac{\partial g}{\partial \underline{\omega}} = \left(\begin{array}{cccc} \frac{1}{R_x}, & -\frac{\tilde{N}_x}{R_x^2}, & 0, & 0 \end{array}\right)^T$, $\frac{\partial h}{\partial \underline{\omega}} = \left(\begin{array}{cccc} 0, & 0, & \frac{1}{R_y}, & -\frac{\tilde{N}_y}{R_y^2} \end{array}\right)^T$,

and

$$\mathrm{cov}(\underline{\omega}) = \begin{pmatrix} \mathrm{var}(\tilde{N}_x) & \mathrm{cov}(\tilde{N}_x, R_x) & \mathrm{cov}(\tilde{N}_x, \tilde{N}_y) & \mathrm{cov}(\tilde{N}_x, R_y) \\ \mathrm{cov}(\tilde{N}_x, R_x) & \mathrm{var}(R_x) & \mathrm{cov}(\tilde{N}_y, R_x) & \mathrm{cov}(R_x, R_y) \\ \mathrm{cov}(\tilde{N}_x, \tilde{N}_y) & \mathrm{cov}(\tilde{N}_y, R_x) & \mathrm{var}(\tilde{N}_y) & \mathrm{cov}(\tilde{N}_y, R_y) \\ \mathrm{cov}(\tilde{N}_x, R_y) & \mathrm{cov}(R_y, R_x) & \mathrm{cov}(\tilde{N}_y, R_y) & \mathrm{var}(R_y) \end{pmatrix},$$

and obtain

$$\begin{aligned}\operatorname{cov}\left(\frac{\tilde{N}_x}{R_x},\frac{\tilde{N}_y}{R_y}\right) &\approx \frac{1}{R_xR_y}\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)-\frac{\tilde{N}_y}{R_xR_y^2}\operatorname{cov}(\tilde{N}_x,R_y)-\\ &- \frac{\tilde{N}_x}{R_x^2R_y}\operatorname{cov}(\tilde{N}_y,R_x)+\frac{\tilde{N}_x\tilde{N}_y}{R_x^2R_y^2}\operatorname{cov}(R_x,R_y). \end{aligned} \tag{6}$$

When we substitute (5) and (6) in (3), knowing that the covariance matrices are symmetric, we obtain

$$\operatorname{var}(N^*)\approx\sum_{x,y=0}^{age_{max}}\left\{\frac{1}{R_xR_y}\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)-\frac{2\tilde{N}_y}{R_xR_y^2}\operatorname{cov}(\tilde{N}_x,R_y)+\frac{\tilde{N}_x\tilde{N}_y}{R_x^2R_y^2}\operatorname{cov}(R_x,R_y)\right\}. \tag{7}$$

Notice that when $x=y$, $\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)=\operatorname{var}(\tilde{N}_x)$ and $\operatorname{cov}(R_x,R_y)=\operatorname{var}(R_x)$ and the summands in (7) coincide with (5).

Equation (7) is somehow problematic to apply, since $\operatorname{cov}(\tilde{N}_x,R_y)$ does not have an explicit analytical formulation. We would like to show that (7) has an upper bound given by

$$\operatorname{var}_2(N^*)=\sum_{x=0}^{age_{max}}\frac{1}{R_x^2}\operatorname{var}(\tilde{N}_x)+\sum_{x,y=0}^{age_{max}}\frac{\tilde{N}_x\tilde{N}_y}{R_x^2R_y^2}\operatorname{cov}(R_x,R_y), \tag{8}$$

the latter being easier to compute (see section 5). Hence we need to show that $\operatorname{var}(N^*)\leq\operatorname{var}_2(N^*)$, that is

$$\begin{aligned}\operatorname{var}(N^*)-\operatorname{var}_2(N^*) &= \sum_{x,y=0}^{age_{max}}\frac{1}{R_xR_y}\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)-\sum_{x=0}^{age_{max}}\frac{1}{R_x^2}\operatorname{var}(\tilde{N}_x)-\\ &- 2\sum_{x,y=0}^{age_{max}}\frac{\tilde{N}_y}{R_xR_y^2}\operatorname{cov}(\tilde{N}_x,R_y)\leq 0. \end{aligned} \tag{9}$$

In general covariance matrices are positive definite, therefore

$$\sum_{x,y=0}^{age_{max}}\frac{1}{R_xR_y}\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)\geq 0;$$

$$\sum_{x=1}^{age_{max}}\frac{1}{R_x^2}\operatorname{var}(\tilde{N}_x)\geq 0;$$

$$-2\sum_{x,y=0}^{age_{max}}\frac{\tilde{N}_y}{R_xR_y^2}\operatorname{cov}(\tilde{N}_x,R_y)\leq 0.$$

Simulations (see next section) show that

$$\sum_{x,y=0}^{age_{max}}\frac{1}{R_xR_y}\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)-\sum_{x=0}^{age_{max}}\frac{1}{R_x^2}\operatorname{var}(\tilde{N}_x)=\sum_{x\neq y;x,y=0}^{age_{max}}\frac{1}{R_xR_y}\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)\leq 0, \tag{10}$$

that is a matrix made of the off-diagonal elements of the matrix $\operatorname{cov}(\tilde{N}_x,\tilde{N}_y)$ is negative definite. In that case (9) holds.

4 A simulation study

A simulation study has been implemented in Gigli *et al.*(2006) in order to verify the assumption $\frac{\tilde{N}_x}{R_x^3}\text{cov}(\tilde{N}_x, R_x) = 0$ in the estimation of the variance of complete prevalence at fixed age (5). Below is a brief description of the simulation; the same results can be used to estimate $\text{cov}(\tilde{N}_x, \tilde{N}_y)$ and $\text{cov}(\tilde{N}_x, \tilde{R}_y)$.

Let W denote the set of SEER-9 patients diagnosed in the period 1975–2001, and Z be the subset of patients diagnosed in the period 1986–2000. For each cancer site of interest let W^+ be the set of patients randomly sampled from W and Z^+ the subset of W^+ containing only patients diagnosed in the period 1986-2000. Notice that whilst W and W^+ are of equal size, the size of Z^+ varies, and in general will differ from the size of Z.

The resampling $W^+ \sim W$ is performed $B = 400$ times and a set of $Z_1^+, \ldots, Z_B^+$ is selected. For each sample Z_b^+ we compute:

a) the 15-year limited duration prevalence $N_{x,b}^+$, for $x = 0, \ldots, age_{max}$, and its standard deviation by SEER*Stat software, and obtain a $[B \times age_{max}]$ matrix, where each row corresponds to a simulation of limited duration prevalence computed for each age group;

b) the incidence and survival parameter vector $\underline{\psi}_b^+$, and obtain a $[B \times p]$ matrix;

c) the completeness index $R_{x,b}^+ = R_x(L; \underline{\psi}_b^+)$, for $x = 0, \ldots, age_{max}$, and its standard deviation via COMPREV software, and obtain a $[B \times age_{max}]$ matrix.

From this data we can compute $\text{cov}^+(\tilde{N}_x, R_y)$ and $\text{cov}^+(\tilde{N}_x, N_y)$ which are plugged in (7):

$$\text{cov}^+(\tilde{N}_x, R_y) = \frac{1}{B}\sum_{b=1}^{B}(N_{x,b}^+ - \bar{N}_x^+)(R_{y,b}^+ - \bar{R}_y^+) \ ,$$

$$\text{cov}^+(\tilde{N}_x, \tilde{N}_y) = \frac{1}{B}\sum_{b=1}^{B}(N_{x,b}^+ - \bar{N}_x^+)(N_{y,b}^+ - \bar{N}_y^+) \ , \tag{11}$$

where $\bar{N}_x^+$ and $\bar{R}_y^+$ are the averages over the B samples of the simulated limited duration prevalence at age x and completeness index at age y, respectively.

Table 1 reports the results of the simulation for three cancer sites: breast and cervix for females, and colon rectum for males; N^* is the estimated complete prevalence obtained from (1); "full" is the variance of complete prevalence as obtained by applying (7); "approx." is the simplified form of the variance of complete prevalence as obtained by applying (8); and finally "naive" is the estimation of $\text{var}(N^*)$ in the hypothesis that all covariances in (3) are null

$$\sum_{x=0}^{age_{max}} \text{var}(N_x^*) \tag{12}$$

The results of the simulation empirically confirm (9): $\text{var}_2(N^*)$ is an easier-to-compute upper bound of $\text{var}(N^*)$.

Table 1: Complete prevalence and standard deviation computed with formulae (7), (8) and (12). Simulated data on female breast and cervix and male colon cancers collected by SEER-9 registry in period 1986-2000; prevalence date: Jan 1, 2001.

		standard deviation		
cancer site	N^*	full: from (7)	approx.: from (8)	naive: from (12)
breast	196,024	416	606	333
cervix	24,116	343	484	113
colon	44,741	202	265	103

5 Computing $\mathrm{cov}(R_x, R_y)$

Recall that in Gigli (2001) variance of the completeness index R_x was computed via the approximation (2):

$$\begin{aligned} \mathrm{var}[R_x(L;\hat{\psi})] &\approx \left(\frac{\partial R_x}{\partial \psi}\Big|_{\psi=\hat{\psi}}\right)^T \hat{V} \left(\frac{\partial R_x}{\partial \psi}\Big|_{\psi=\hat{\psi}}\right) \\ &= \sum_{i,j=1}^{p} \left(\frac{\partial R_x}{\partial \psi_i}\right) \times \left(\frac{\partial R_x}{\partial \psi_j}\right) \times \mathrm{cov}(\psi_i, \psi_j) \end{aligned} \tag{13}$$

where $\partial R_x/\partial\psi = (\partial R_x/\partial\psi_1, \ldots, \partial R_x/\partial\psi_p)$ is the vector of partial derivatives of R_x with respect to the incidence and survival parameters and $\hat{V}$ is the estimated covariance matrix of $\hat{\psi}$.

Here (4) applies

$$\mathrm{cov}(R_x, R_y) \approx \sum_{i,j=1}^{p} \left(\frac{\partial R_x}{\partial \psi_i}\right) \times \left(\frac{\partial R_y}{\partial \psi_j}\right) \times \mathrm{cov}(\psi_i, \psi_j). \tag{14}$$

A Fortran program called "comprev.f" was developed in order to compute (13):

```
c computing the s.e. of the completeness index
write(3,*)
write(3,*) ' STANDARD ERROR OF COMPLETENESS INDEX'
write(3,*) 'age completeness se(completeness)'
do 3014 jage=1,nage
3014 var(jage)=0.0d0
do 3215 jage=1,nage
agecor=17+jage*5
do 3115 ider=1,id
do 3125 kder=1,id
3125 var(jage)=var(jage)+ dercompl(jage,ider)*dercompl(jage,kder)*cov(ider,kder)
3115 continue
var(jage)=dsqrt(var(jage))
write(3,500) agecl(jage),compl(jage),var(jage)
write(4,700) agecor,compl(jage),var(jage)
3215 continue
500 format(A5,1X,f6.4,1x,f14.12)
```

700 format(i3,1x,f6.4,1x,f14.12)

Now, in order to compute (14) the following modification needs to be implemented:

```
do 3215 jage=1,nage
do 3220 lage=1,nage
agecor=17+jage*5
do 3115 ider=1,id
do 3125 kder=1,id
3125 var(jage,lage) = var(jage,lage) + dercompl(jage,ider) * dercompl(lage,kder) * cov(ider,kder)
3115 continue
3220 continue
3215 continue
```

References

- Capocaccia R, De Angelis R (1997). Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine*, **16**, 425–440.
- Clegg LX, Gail MH, Feuer EJ (2002). Estimating the variance of disease-prevalence estimates from population-based registries. *Biometrics*, **55**, 1137–1144.
- Gail MH, Kessler L, Midthune D, Scoppa S (1999). Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. *Biometrics*, **55**, 1137–1144.
- Gigli A. (2001) The variance of the completeness index, IRP W.P. 3/2001.
- Gigli A, Mariotto A, Clegg LX, Tavilla A, Corazziari I, Hachey M, Scoppa S, Capocaccia R (2006). Estimating the variance of cancer prevalence from population-based registries. *Statistical Methods for Medical Research*, **15**, p. 235–253.
- Verdecchia A, Capocaccia R, Egidi V, Golini A (1989). A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine,* **8**, *201–216.*