RESEARCH PAPER

An application of the two-source capture–recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy

E Crocetti, G Miccinesi, E Paci, M Zappa

(Received 2 April 2001; accepted 14 May 2001)

The purpose of this study was to apply a two-way capture–recapture method to estimate the Tuscany Cancer Registry completeness, taking into account the presence of dependence between sources. Cases incident during 1995–1996 were flagged according to three sources of information: clinical notes, pathological reports and death certificates. For each group of cases notified by one source the dependence between the other two has been quantified and the completeness has been estimated by a two-way capture–recapture method. When only two (or substantially two) sources are dependent on each other it is possible to correct for it by pooling the two sources in a single group and comparing it with the remainder source by a two-way capture–recapture method. The capture–recapture method has been applied to the overall 12 387 incident cases and to 1569 female breast and 1443 lung cancer cases. After correction for the greatest dependence among the three couples of sources of information, the estimates of completeness were 97.4% for the whole case series, 88.7% for female breast and 99.6% for lung cancer. With the limit of multiple strong dependence between sources, the two-way capture–recapture method seems a simple and useful tool for estimating the completeness of cancer registration.

Key words: Cancer registry, capture–recapture methods, completeness.

Introduction

Population-based cancer registries are designed to measure cancer incidence rates arising from a defined population, avoiding case selection and incompleteness. However, complete case ascertainment is difficult to achieve. The traditional estimate of under-ascertainment is based on the proportion of cases known from the death certificate only (DCO). This method is widely used because it is straightforward and based on death certificates, which represent one of the basic information sources of any cancer registry (Parkin *et al.*, 1997). In recent years there has been a growing interest in capture–recapture (CR) methods (Robles *et al.*, 1988; Brenner *et al.*, 1994; Brenner *et al.*, 1995; Shouten *et al.*, 1994; Dockerty *et al.*, 1997; Seddon and Williams, 1997). The CR method has been originally developed and widely applied in wildlife science as a tool to estimate the size of free-living animal populations (Cormack, 1968). Briefly, samples of animals are captured, tagged, released and recaptured. The number of animals recaptured and newly captured in each sample allows the estimate of the population

UO Epidemiologia Clinica e Descrittiva, CSPO, Via di San Salvi 12, 50135 Florence, Italy. Correspondence to: E Crocetti. Fax: (+39) 055 679954. E-mail: e.crocetti@cspo.it

0959-8278 © 2001 Lippincott Williams & Wilkins

size. Similarly, the estimate of cancer burden in a population, by a cancer registry, is based on the combination of case ascertainment by multiple incomplete sources.

The fundamental assumption of the CR method is that each capture is independent from the others, which means that the probability of being captured is not modified by the result of a previous capture. This is improbable for cancer patients. For example, pathological reports are more frequent for hospitalized patients, leading to positive dependence between these two sources. Besides, being successfully treated reduces the probability of being notified by death certificate, leading to negative dependence. In the case of positive dependence between two sources, CR methods underestimate the true number of cases and overestimate completeness, whereas negative dependence overestimates the true number of cases, leading to underestimation of completeness (Brenner, 1995).

The aim of this paper is to evaluate the basic two-source CR method as a tool to estimate the completeness of the Tuscany Cancer Registry. When positive or negative dependence between sources is present, possible adjustments for the application of CR method are outlined.

Material and methods

In the provinces of Florence and Prato, central Italy (about 1 200 000 inhabitants), the Tuscany Cancer Registry (RTT), a population-based cancer registry, has been active since 1984. The description of the criteria for collection, registration and analysis followed by the registry has been presented elsewhere (Parkin *et al.*, 1997). Briefly, the registry receives clinical notes from all the public hospitals of the Tuscan region and from those of the private ones that are financed by the Regional Health Authority; clinical notes are also received from the main Italian oncological hospitals and, when necessary, general practitioners are involved. Pathological reports are collected from all the pathological departments active in the registry area and from the main ones of the region. Death certificates of resident subjects are retrieved from the Regional Mortality Registry.

In the analysis, 12 387 cancer patients diagnosed between 1995 and 1996 in the provinces of Florence and Prato were included, whereas multiple primary cancers (1270 cases) and non-melanomatous skin cancers (1166 cases) were excluded.

For each patient the documentation used to define the incident case has been flagged as:

(a) *Clinical*. If the registration was based on (or also on) hospital discharge for cancer or was based on information received from the general practitioner.

(b) *Pathological*. If registration was based on (or also on) cyto/histological report positive or suspect for cancer. Eight cases diagnosed on the basis of autopsy (performed in pathological departments) were included in this group.

(c) *Death certificate*. If registration was based on (or also on) death certificate reporting a tumour death cause (ICD-9 140–239). At the time data were analysed, 1997 death certificates were available.

In simple algebraic terms, as shown in Table 1, among cases known since reported by source z it is possible to evaluate the individual contribution of the source y as $(a + b)$ and source x as $(a + c)$. Cases known by traditional method would be $(a + b + c)$, without any estimate of the quantity d. Using the CR method it is possible, by a maximum likelihood method, to estimate d as $[(b \times c)/a]$ and N as $[(a + b) \times (a + c)/a]$ (Bishop *et al.*, 1975).

This method assumes the independence between the two sources, but this is not usually the case.

To evaluate the positive or negative dependence between sources and its effect on completeness estimate, we used the method proposed by Brenner (1995). First, we define a subset of cases as those identified by one of the three main sources, and then the completeness of registration due to the two other sources is compared with the corresponding two-source CR method.

We can estimate the positive dependence as $\{a/N - [(a + b)/N \times (a + c)/N]\}$, and the maximum positive dependence as $\{(a + c)/N \times [1 - (a + b)/N]\}$ if $[(a + c)/N < (a + b)/N]$ (otherwise factors should be inverted). It is therefore possible to indicate the positive dependence between the two sources as the ratio between the previous two formulas (Brenner, 1995). The negative dependence is estimated as $\{[(a + b)/N \times (a + c)/N] - a/N\}$; the maximum negative dependence as $\{1 - [(a + c)/N] \times (1 - (a + b)/N\}$, and the proportion of maximum negative dependence due to the dependence among the two analysed sources as the ratio between the previous two formulas (Brenner, 1995).

If the two analysed sources are independent, the CR method estimates exactly the number of lacking (d) cases. If the two sources are positively dependent, the CR method underestimates d and overestimates the completeness; in contrast, if the sources

Table 1. Distribution of cases reported by source z in relation to the presence or absence of notifications from source x and y

		Source x		
		Yes	No	
Source y	Yes	a	b	$a + b$
	No	c	d	$c + d$
		$a + c$	$b + d$	N

Notified cases by source z.

are negatively dependent d is overestimated and the completeness is underestimated.

The informative contribution of each source, the presence and direction of dependence among three couples of the three main informative sources are computed for all the data set of the RTT and estimates of completeness which correct for such dependence are computed. The probability of being notified by each source may differ according to the cancer site (age, prognosis, clinical course, treatment, histology accessibility, etc.), therefore a further analysis has been carried out for two specific cancer sites: female breast and lung (both sexes).

Results

In Table 2, the individual and combined contribution for the three main information sources is shown. Overall, for 88% of the 12 387 cases there is a clinical notification, for 77% a pathological report and for about 43% the death certificate. A pathological report is available for 94.4% of breast cancer cases and for only 60.7% of lung cancer cases. On the other hand the death certificate is present for 79.7% of lung cancer cases but for only 8.3% of breast cases. For both cancer sites, at least one clinical note is available for over 94% of the cases.

In Table 3, an analysis within the cases identified by each of the three information sources is shown to evaluate the informative contribution of the other two.

Among cases known by pathological reports (Table 3) we can quantify the dependence between clinical notes and death certificates. For all sites together, there is a slight positive dependence (0.02), which is almost half (45%) of the maximum possible (0.04). This leads to an underestimate of lacking cases corresponding to an overestimate of completeness. In fact, the CR method leads to an estimate of 8957 cases, sensibly lower than the 'true' 9555 (all cases known by pathological reports). The underestimate of the true value is less severe than the one that would be derived from the traditional method (8567, 10.4% less than 9555).

For breast cancer (Table 3) among the 1480 cases known from the pathological report, there is a slight negative dependence (1.6% of the maximum possible) between clinical reports and death certificates, resulting in a slight overestimation of incompleteness (1498 versus 1480 cases notified by pathological reports), the estimate from traditional method is further from the true value (1418 versus 1480).

For lung cancer there is a small positive dependence between clinical and death sources and therefore a slight underestimate of lacking cases (873 versus 876). When the two sources clinical notes and death certificates are grouped together and compared with pathological reports in a two-way CR method, as shown in Table 3, we eliminate the effect of the dependence between the grouped sources. For the whole series the estimate, which corrects for a strong dependence between clinical notes and death certificates, leads to 12 715 total cases corresponding to a RTT completeness of 97.4%. The corresponding values for female breast and lung cancer are 99.7% and 99.4% respectively.

Between pathological reports and death certificates (among cases known throughout clinical notes), there is negative dependence for all sites, female breast and lung cancer. For all the sites together, the value of negative dependence is not very high (0.059) but it is almost half of all the possible negative dependence (0.44), leading to a strong overestimate of cases (13 237 versus 10 944). Also for breast cancer the negative dependence between pathological reports and death certificates is quite high and

Table 2. Tuscany Cancer Registry 1995–1996: number of known cases of lung, female breast and all sites cancer, with the percentage of contribution of each information source and their combinations

Cancer site	*N* (%)	Clinical note only	Pathological report only	Death certificate only	Clinical + pathological + death	Pathological + death	Clinic + pathological	Clinic + death
Lung	1 443 (100)	91 (6.3)	13 (0.9)	31 (2.2)	639 (44.3)	35 (2.4)	189 (13.1)	445 (30.8)
Female breast	1 569 (100)	46 (2.9)	62 (4.0)	16 (1.0)	83 (5.3)	5 (0.3)	1330 (84.8)	27 (1.7)
Any site	12 387 (100)	809 (6.5)	992 (8.0)	222 (1.8)	3061 (24.7)	229 (1.9)	5273 (42.6)	1801 (14.5)

Table 3. Tuscan Cancer Registry 1995–1996. Dependence between two sources of information in each group of cases identified by the third one and results from capture–recapture two-way method corrected for dependence

Cases known from cancer site	True (n)	Known (n)	Estimate (n)	Dependence	% of max dependence	Known	Estimate	%
	Pathological report					Pathology versus clinical and death certificate completeness		
ALL	9 555	8 567	8 957	Positive	45	12 387	12 715	97.4
Breast	1 480	1 418	1 498	Negative	2	1 569	1 573	99.7
Lung	876	863	873	Positive	5	1 443	1 452	99.4
	Clinical notes					Clinical versus pathology and death certificate completeness		
ALL	10 944	10 135	13 237	Negative	44	12 387	12 502	99.1
Breast	1 486	1 440	1 873	Negative	32	1 569	1 572	99.8
Lung	1 364	1 273	1 405	Negative	18	1 443	1 449	99.6
	Death certificate					Death certificate versus clinical and pathology completeness		
ALL	5 313	5 091	5 226	Positive	17	12 387	12 695	97.6
Breast	131	110	117	Positive	64	1 569	1 769	88.7
Lung	1 150	1 119	1 143	Positive	9	1 443	1 451	99.4

For each subset of cases and each cancer site the following values are shown: the true number of cases (those defined by the notifying source), the number of cases known by the other two sources and the number of cases estimated by the two-way capture–recapture method. The dependence between the two sources analysed in each subset is shown with its value as the percentage of the maximum possible one. In the last two columns the number of cases estimated by a two-source capture–recapture method based on the group defining source and the other two together and the corresponding estimate of completeness percentage are shown.

the overestimate is significant (1873 versus 1486). For lung cancer, a lower value of the negative dependence still corresponds to an overestimate but is lower than in the previous two situations. Adding together the pathological source and the death certificates and comparing them with the clinical reports in a two-source CR model we have an estimate of completeness which takes account of the dependence between pathology and death certificates. The estimates of completeness, shown in Table 3, are 99.1% for all the registry's series, 99.8% for breast cancer and 99.6% for lung cancer.

Finally, when we consider those cases identified by death certificate, we can compare clinical and pathological sources with a two-way CR method (Table 3). As expected, between these two sources there is positive dependence for all the three cancer series analysed. It is of medium level for all the sites together (17% of the possible maximum), very high for breast cancer (64% of the maximum) and quite slight for lung cancer (9%), with corresponding levels of overestimation of completeness. The two-way CR method, which takes account of the dependence between clinical notes and pathological reports, leads to completeness estimates of 97.6% for the whole RTT case series, 88.7% for female breast cancer and 99.4% for lung cancer.

Discussion

The present study confirmed that the CR method is, in its simplest two-way form, a straightforward tool that can be applied easily to cancer registry original series.

The application of the CR method in human studies has been hampered by the lack of independence between different sources. We confirmed such dependence between the three main sources of information in a cancer registry: clinical notes, pathological reports and death certificates. For example, a strong negative dependence was found between pathological report and death certificate in all the three situations analysed. It is possible that cases diagnosed in an advanced stage, with a poor prognosis, did not undergo surgical interventions − thus the proportion of death certificates is higher when that of pathological reports is lower. It is worth stressing that the relevance of the death certificate as an informative source is related to the cancer-specific lethality but also to the length of follow-up availability, which may vary in different registries. On the other hand, a positive dependence has been seen between clinical notes and pathological reports. Cytohistological examinations are more frequently performed during an hospital admission, leading to this positive dependence.

We quantified the effect of dependence between sources and compared the estimates obtained by CR method with results from the traditional method. In most cases, the CR method offered a better result than the traditional one. Also without any adjustment, the CR method may offer useful estimates; in the case of positive dependence the estimate of missed cases may be considered the lower limit of the true number and in the case of negative dependence it may be considered the upper limit.

We applied the method proposed by Brenner (1995) by using a two-source CR method 'adjusting' for the dependence. If dependence exists between only two of three sources (or especially between two) it is possible to 'correct' for its effect by considering the two most dependent sources together and applying the CR method between these two as a single group and the other source. For example, for lung cancer the most evident dependence was the negative one between pathological reports and death certificates. The CR estimate of completeness computed between these two sources together and the clinical notes should give the best estimate among the three possible ones, 99.6%.

The percentage of incompleteness estimated by the CR method for the overall RTT series varies from 0.9% to 2.4%. This value agrees with the percentage of cases known from the death certificate only (DCO 1.8%). As DCO are strictly related to case fatality, the DCO rate is expected to be a priori lower for non-fatal cases and higher for fatal ones; in fact, the percentage of DCO was 1 for female breast cancer and 2.2 for lung cancer.

If the sources are all three heavily dependent, two by two, it is not possible to adjust by just coupling two of these. For example, for female breast cancer when we add pathological report with clinical notes we correct for the positive dependence between these two sources but not for the negative dependence between death certificates and pathological reports. The estimates derived from CR methods lead to a wide range of lost cases (from 3 to 20) with an unreliable estimate of completeness (from 88.7% and 99.8%). However, when a low informative source (as death certificates that are available for breast cancer for only 8.3% of the cases) is combined with a strong negative dependence as that between death certificates and pathological reports (for breast cancer cases known from clinical notes), there is a very high overestimate of lacking cases (Brenner, 1995). Therefore, for breast cancer the best estimate of completeness should be the one which compares, by a two-way CR method, the death certificates and the pathological reports grouped together with the clinical notes; this estimate leads to a reasonable number of missed cases (3) and to 99.8% of completeness.

In situations where the sources are all dependent, two by two, a regression log-linear modelling, which includes all sources together, is needed (Robles *et al.*, 1988; Brenner *et al.*, 1994; Shouten *et al.*, 1994; Hook and Regal, 2000). When three sources are considered the all-two-way-interactions seem to perform optimally (Hook and Regal, 2000). However, if the three sources are all dependent of each other, also with the modelling approach the three-way interaction term remains unknown, unless a further source which estimates those cases not identified by none of the three sources is included (Shouten *et al.*, 1994). The modelling approach is also necessary when we want to evaluate other variables, which may be related to notification sources, such as age (Robles *et al.*, 1988; Brenner *et al.*, 1994; Shouten *et al.*, 1994; Dockerty *et al.*, 1997).

The study analysed some cancer sites but the same method can be applied to all the other sites. In fact, as stated above, it is possible to identify the highest dependence among couples of sources and 'adjust' for it. This could be applied, for example, to the positive dependence between clinical notes and death certificates for cervix uteri cancer (57% of the maximum positive possible) or for ovary cancer (36%). We could also 'adjust' for the negative dependence between histology and clinical reports for prostate cancer (44% of the maximum negative possible).

For cancer registries the pathology report is considered to be the gold standard, and no further efforts are usually carried out to discover some other missed sources of information when the pathology has been collected. Therefore, some of the patients with pathology only may also have a hospital admission and/or the death certificate missed due to errors in one of the keys used for the linkage (name, surname, date of birth) between data sets. In addition, some of the patients admitted to hospital during 1996 may not be discharged before the end of the year.

False positive notifications due to errors in the digits for cancers do not matter because the registry has a policy which requests more than one source of information to define a cancer case, except for pathology for which the diagnosis is explicit, and for death certificate only. Therefore, the inability of CR methods to give any information on overcapture does not seem important, at least for the RTT. False negative notifications cannot be avoided and affect both incidence rates and CR methods; the latter are indeed used to evaluate the registry's completeness. Some cases with all their clinical history in foreign countries may be missed (Shouten *et al.*, 1994). Finally, in the Tuscan Region some private clinics have no financial relationship with the Public Health Authority and therefore do not take part in its information flow. Cases with only this information may be lost. Six per cent of the overall regional hospital admissions occur in 21 private clinics, of which five are completely private. Therefore, the average information lost should not be over about 1%, if neither pathological report nor death certificate are available.

In conclusion, the application of the two-source CR method on the case series of the Tuscany Cancer Registry has shown that it may be a useful tool for estimating the completeness of case ascertainment. The dependence present among the information sources used by a cancer registry may be measured and it does not hamper the CR method application if it is present only (or substantially) between two of three information sources.

References

Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. **[Q4]**

Brenner H (1995). Use and limitations of the capture–recapture method in disease monitoring with two dependent sources. *Epidemiology* **6:** 42–8.

Brenner H, Stegmaier C, Ziegler H (1994). Estimating completeness of cancer registration in Saarland/Germany with capture–recapture methods. *Eur J Cancer* **30:** 1659–63.

Brenner H, Stegmaier C, Ziegler H (1995). Estimating completeness of cancer registration: an empirical evaluation of the two source capture–recapture approach in Germany. *J Epidemiol Commun Health* **49:** 426–30.

Cormack RM (1968). The statistics of capture–recapture methods. *Oceanogr Mar Biol Annu Rev* **6:** 455–506.

Dockerty JD, Becroft DMO, Lewis ME, Williams SM (1997). The accuracy and completeness of childhood cancer registration in New Zealand. *Cancer Causes Control* **8:** 857–64.

Hook EB, Regal RR (2000). Accuracy of alternative approaches to capture–recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* **8:** 771–9.

Parkin DM, Muir CS, Whelan SL, *et al*. (1997). *Cancer Incidence in Five Continents*, Vol VII. IARC Scientific Publications No. 143. IARC, Lyon.

Robles SC, Marrett LD, Clarke EA, Risch HA (1988). An application of capture–recapture methods to the estimation of completeness of cancer registration. *J Clin Epidemiol* **41:** 495–501.

Seddon DJ, Williams EMI (1997). Data quality in population-based cancer registration: an assessment of the Merseyside and Cheshire Cancer Registry. *Br J Cancer* **76:** 667–674.

Shouten LJ, Straatman H, Kiemeney ALM, Gimbrère CHF, Verbeek ALM (1994). The capture–recapture method for estimation of cancer registry completeness: a useful tool? *Int J Epidemiol* **23:** 1111–16.