International Journal of Epidemiology
© International Epidemiological Association 1994

Vol. 23, No. 6
Printed in Great Britain

# The Capture-Recapture Method for Estimation of Cancer Registry Completeness: A Useful Tool?

LEO J SCHOUTEN,*,**,† HUUB STRAATMAN,* LAMBERTUS A L M KIEMENEY,†,‡ CHARLES H F GIMBRÈRE§ AND ANDRÉ L M VERBEEK*,†

Schouten L J (Department of Medical Informatics and Epidemiology, University of Nijmegen, PO Box 9101, NL-6500 HB Nijmegen, The Netherlands), Straatman H, Kiemeney L A L M, Gimbrère C H F and Verbeek A L M. The capture-recapture method for estimation of cancer registry completeness: A useful tool? *International Journal of Epidemiology*, 1994; **23**: 1111–1116.

*Background.* In this paper we investigated whether the capture-recapture method is useful for a cancer registry to monitor its completeness of case ascertainment on a routine basis.

*Methods.* The capture-recapture method was used to estimate the completeness of case ascertainment in three regional cancer registries in the Netherlands, which are based on case finding by pathology laboratories and hospitals.

*Results.* Completeness was estimated to be 98.3%. The estimate of completeness was dependent on age and cancer site, with lower estimates of completeness for skin cancer and lymphatic and haematopoietic malignancies and for the age group ≥75 years.

*Conclusions.* A major drawback of the capture-recapture method is its inability to estimate the number of cases that are not routinely notified to the registry by one or both notification sources. Another limitation is the lack of statistical power to detect incompleteness in an early stage. It is concluded that the capture-recapture method is not useful for everyday surveillance of completeness in cancer registration.

Completeness of cancer registration is important because (selective) incomplete case ascertainment may lead to misinterpretation of trends in time and place. It can also cause bias in epidemiological research based on the registry. Several methods have been proposed for testing completeness.[1] The most reliable method is independent case ascertainment, which involves linkage of individual records from the cancer registry with the records of an independent survey in the same catchment area and time.[1] This, however, requires considerable effort and it is not feasible to apply the method on a routine basis.

In zoology a method has been developed to estimate the size of wildlife populations, i.e. the capture-recapture method. In different samples animals are counted and tagged. By counting the proportions of animals that have been tagged in other samples, the size of the total population can be estimated. This method has also been used in epidemiological studies to estimate the prevalence of a disease.[2–4] Recently, the capture-recapture method was used to estimate the completeness of a cancer registry.[5]

In this study we investigated whether the capture-recapture method is useful for a cancer registry to monitor its completeness of case ascertainment on a routine basis. For this study the 1990 data from three regional cancer registries in the Netherlands were available for analysis.

## METHOD

### *The Cancer Registries*

The regional cancer registries of the Comprehensive Cancer Centres, IKL (located in Maastricht), IKMN (Utrecht) and IKO (Nijmegen) were established in 1984, 1985 and 1986, respectively. These three registries cover a total population of 3.3 million inhabitants. From 1989 onwards, all hospitals and pathology laboratories in the regions participated in (one of) these registries.

The cancer registries receive lists of newly diagnosed cancer cases on a weekly basis from the pathology departments in the regions. In addition, lists of discharge diagnoses of hospitalized cancer patients are obtained

* Department of Medical Informatics and Epidemiology, University of Nijmegen, PO Box 9101, NL-6500 HB Nijmegen, The Netherlands.
** Department of Cancer Registration, Comprehensive Cancer Centre IKL, Maastricht, The Netherlands.
† Department of Cancer Registration, Comprehensive Cancer Centre IKO, Nijmegen, The Netherlands.
‡ Dutch Cancer Society, The Netherlands.
§ Department of Cancer Registration, Comprehensive Cancer Centre IKMN, Utrecht, The Netherlands.

from the hospitals. These reporting procedures are computerized for the greater part. Following these notifications, the medical records of newly diagnosed patients (or tumours) are collected and the relevant information for the cancer registry is abstracted from the charts by trained registry personnel. All malignancies, including non-invasive malignancies, are recorded, with the exception of basal cell carcinomas of the skin and non-invasive cervical cancer.

From January 1990 onwards, the sources of notification were recorded for each tumour record. Possible sources were pathology department (PA), hospital medical record department (MR) and other sources (OS), e.g. radiotherapy department. Because the number of records reported by OS only was very small (N=30), such records were excluded for this study. In addition, only a patient's first malignancy was selected for analysis. Due to shortcomings in the automatic processing of the MR notifications, the MR source for second (and subsequent) malignancies was not accurate.

*The Capture-Recapture Method*

In 1990, in the catchment areas of the regional cancer registries IKL, IKMN and IKO an unknown number of patients were diagnosed with a first primary malignancy. Every patient had a certain probability of being entered into the pathology registration system (PA) and a certain probability of being entered into the hospital discharge system (MR). The cancer registries combined both data sources in order to achieve complete ascertainment. However, for a number of reasons, some patients were registered neither in PA nor in MR. Therefore, instead of cancer incidence N, an incomplete number of cases, T was registered by the registries.

*Simple analysis.* In order to estimate the completeness of ascertainment by the cancer registries the capture-recapture method, as described in Bishop *et al.*[6] was applied. With this method, all records in the cancer registries are classified according to source of notification: $PA^+MR^+$, $PA^+MR^-$ and $PA^-MR^+$, respectively. If notification by PA does not influence the chance of being notified by MR (and vice versa), the number in the $PA^+MR^+$ cell is hypergeometrically distributed, conditional on the number of records captured by each of the notification sources. Consequently, the total number of eligible cases can be estimated by ((a+b)*(a+c))/a and the number of cases not registered can be estimated by b*c/a (where a stands for the number of records with $PA^+MR^+$ notification, b for $PA^+MR^-$ notification and c for $PA^-MR^+$ notification).

*Modelling.* The proportion of completeness can be estimated for different sites, ages etc. In that case, it may be more efficient to use a modelling approach. With such an approach, a log-linear regression model is chosen which adequately represents the observed data with one or two of the notification sources. The choice of the model is based on goodness of fit, including the distributions of residuals and the residual deviance (−2 log-likelihood for Poisson probability distributions). The two-way interaction terms were chosen from all the remaining two-way interaction terms, not included in the current model, with the highest change in deviance divided by degrees of freedom. If the new model was significantly better than the current model, the interaction term was added to the model. If the model fit was good, we found a satisfactory model, otherwise we had to repeat this method to look for an additional two-way (or three-way) interaction term. When we finally found a satisfactory model, redundant two-way interaction terms were deleted. After model selection, the resulting parameter estimates and interaction terms can be used to estimate the stratum-specific numbers of unregistered cases with cancer.

Comparable to the more simple analysis mentioned above, in the modelling approach it is assumed that the data sources are independent. It is possible to include a cross-product interaction term for the data sources in the model, but the parameter for such a term will be aligned with zero because there are no observations with $PA^-MR^-$. For formulae calculating confidence intervals for both the simple method and the modelling approach we refer to Robles *et al.*[5]

In the analyses, both methods were used to estimate the completeness of the cancer registry. In the simple method the number of missed cases was calculated by summation of the number of missed cases per stratum, because of the phenomenon that the capture-recapture estimates can vary in subgroups because of 'variable catchability'.[7] In the log-linear regression analyses, the observed data were modelled as a function of data source (MR and/or PA), primary site, age (<65, 65 –<75 and ≥75 years) and cancer registry (IKL versus IKMN versus IKO). According to anatomical localization, the primary sites were combined into 11 groups. Site, age group and registry-specific strata with an empty cell (no records with only-PA or only-MR notification) were excluded, because the zero from the empty cell could not adequately be modelled.

For model fitting, the statistical program GLIM was used.[8]

## RESULTS

In total, the regional cancer registries IKL, IKMN and IKO recorded 12 570 primary malignancies which were diagnosed in 1990. Of these, 1957 were excluded

TABLE 1 *Distribution of first primary malignancies diagnosed in 1990 according to cancer registry and source of notification*

| Regional cancer registry | MR[a] only | | PA[b] only | | MR[a] and PA[b] | | Total |
|---|---|---|---|---|---|---|---|
| | T[c] | (%) | T[c] | (%) | T[c] | (%) | T[c] |
| IKL | 217 | (7.6) | 387 | (13.6) | 2248 | (78.8) | 2852 |
| IKMN | 221 | (6.1) | 597 | (16.5) | 2803 | (77.4) | 3621 |
| IKO | 371 | (9.0) | 689 | (16.6) | 3080 | (74.4) | 4140 |
| Total | 809 | (7.6) | 1673 | (15 8) | 8131 | (76.6) | 10 613 |

[a] MR = Notification by a medical records department of a hospital.
[b] PA = Notification by a pathology department.
[c] T = Number of recorded malignancies.

TABLE 2 *Estimated completeness according to site, age group and cancer registry*

| Category | Total recorded T | % completeness (95% confidence interval) | | | |
|---|---|---|---|---|---|
| | | Simple analysis | | Modelling approach | |
| Site | | | | | |
| Oral cavity and upper respiratory tract | 224 | 99.2 | (97.9–100) | 99.3 | (98.6–100) |
| Gastrointestinal tract | 2037 | 99.3 | (99.0–99.7) | 99.3 | (99 1–99.6) |
| Deep-seated digestive organs | 416 | 97 5 | (95.6–99.5) | 97.5 | (95.5–99.5) |
| Bronchus, lung and pleura | 1686 | 98.1 | (97.4–98.9) | 98.1 | (97.6–98.8) |
| Skin | 478 | 92.9 | (87.2–99.5) | 93.0 | (87.0–99.9) |
| Breast | 1687 | 99.7 | (99.4–100) | 99.7 | (99.6–99.9) |
| Female genital organs | 746 | 99.3 | (98.7–100) | 99 4 | (99.0–99.8) |
| Male genital organs | 863 | 98.6 | (97.7–99.6) | 98.6 | (98 0–99.2) |
| Urinary tract | 963 | 99.4 | (98.9–100) | 99.4 | (99.1–99.7) |
| Lymphatic/haematopoietic tissue | 706 | 95.2 | (93.3–97.3) | 95.3 | (93.5–97.1) |
| Other sites and unknown primary | 807 | 96.4 | (94.8–98.1) | 96.5 | (95.2–97.9) |
| Age groups | | | | | |
| <65 years | 4655 | 98.9 | (98.5–99 3) | 98.9 | (98.5–99.3) |
| 65–<75 years | 3010 | 98.7 | (98.3–99.2) | 98.7 | (98.3–99 1) |
| ≥75 years | 2948 | 96.9 | (95.7–98.1) | 96.9 | (95.8–98.1) |
| Cancer registry | | | | | |
| IKL | 2852 | 98.6 | (98.1–99.1) | 98.6 | (98 1–99.2) |
| IKMN | 3621 | 98.5 | (97.7–99.2) | 98.5 | (97.8–99.2) |
| IKO | 4140 | 97.9 | (97.1–98.6) | 97.9 | (97.2–98.6) |
| Total | 10 613 | 98.3 | (97.9–98.7) | 98.3 | (97.9–98.7) |

because the malignancy was not the first primary (N = 1078), because some strata had no records reported by PA-only or MR-only (N = 849) or because notification was from a source other than PA and MR (N = 30). Thus, 10 613 records were available for analysis.

In Table 1 the distribution of these malignancies is shown according to regional registry and notification source. In total, 809 malignancies (7.6%) were notified by MR alone, 1673 (15.8%) malignancies by PA alone and 8131 malignancies (76.6%) by both PA and MR.

Using the simple method the number of missed cases was estimated to be 187. Completeness was estimated to be 98.3%. In Table 2 the results are shown, according to site, age group and cancer registry. Completeness varied by site from 92.9% for skin malignancies

TABLE 3 *Choice of model for assessing the completeness of the cancer registry*

| Model[a] | Deviance | d.f.[b] | fit: | Difference from previous model | | Test against previous model |
|---|---|---|---|---|---|---|
| | | | *P* | Deviance | d.f. | *P* |
| Main effects only | | | | | | |
| (1) s + r + a + pa + mr | 2705.9 | 244 | 0.000 | – | – | – |
| Adding 2-way interactions | | | | | | |
| (2) + s.mr | 1560.4 | 234 | 0.000 | −1145.5 | −10 | 0.000 |
| (3) + s.pa | 1060.4 | 224 | 0.000 | −500.0 | −10 | 0.000 |
| (4) + s.a | 527.7 | 205 | 0.000 | −532.7 | −19 | 0.000 |
| (5) + a.pa | 438.3 | 203 | 0.000 | −89.4 | −2 | 0.000 |
| (6) + s.r | 338.1 | 184 | 0.000 | −100.2 | −19 | 0.000 |
| (7) + r.pa | 301.0 | 182 | 0.000 | −37.1 | −2 | 0.000 |
| (8) + a.mr | 269.8 | 180 | 0.000 | −31.1 | −2 | 0.000 |
| Adding 3-way interactions | | | | | | |
| (9) + s.a.mr | 206.8 | 161 | 0.009 | −63.0 | −19 | 0.000 |
| (10) + s.r.mr | 155.8 | 140 | 0.170 | −51.0 | −21 | 0.000 |
| (11) + s.r.pa | 119.7 | 121 | 0.517 | −36.1 | −19 | 0.010 |
| Deleting redundant 2-way interactions, final model s + r + a + pa + mr + a.pa + s.a.mr + s.r.mr + s.r.pa | 119.7 | 121 | 0.517 | – | – | – |

[a] Used terms in model: s = site; r = registry; a = age group; pa = notification by pathology laboratory; mr = notification by medical records department; s.a = interaction factor of site and age group.
[b] d.f. = degrees of freedom.

(excluding basal cell cancer) to 99.7% for cancer of the breast. The registries' completeness was estimated to be somewhat lower for cases >75 years (96.9%).

In the log-linear models, the two sources of notification, age, site of malignancy and cancer registry were entered. The model contained parameters for the 11 site groups, the three cancer registries, the three age groups and the two data sources, as well as one two-way interaction and three three-way interactions (Table 3).

The results using the modelling approach were similar to those from the simple analysis (Table 2). Also, for the various strata the difference in estimates of completeness between the two methods did not exceed 0.1%.

## DISCUSSION

In this study we used capture-recapture methods to estimate the completeness of three regional cancer registries in the Netherlands. The results indicate a very high degree of completeness, estimated at 98.3%.

The estimates of completeness were lowest for malignancies of the skin, lymphatic and haematopoietic tissue and other or unknown sites. These results are as expected. Skin malignancies are treated relatively frequently in outpatient clinics, and will then be reported by the pathology department only. The chance that an error in the notification procedure will lead to a missed case for cancer registration is therefore higher. Haematological malignancies are not always diagnosed by pathologists (but by a haematologist or an internist), which leaves the medical records department as the only notification source for a substantial part of these tumours.

We compared the simple method with a log-linear regression model and found the differences between the results were small which confirmed the results of Robles.[5]

The capture-recapture method is only capable of estimating the number of malignancies that could have been notified by pathology laboratories or hospitals. Although in general the notification procedures were handled automatically, notifications may have been missed. For example, it appeared that

some of the notifications from the medical records departments concerned malignancies with microscopic verification, but for some reason they were not entered into the cancer registry database. This was probably due to administrative errors at the pathology laboratory or the cancer registry itself.

In the Netherlands two studies have estimated the completeness of the cancer registry using the independent case ascertainment method.[9,10] Berkel estimated the proportion of missed cases (known only to the general practitioner) at 1.3%.[9] In a study of the IKL cancer registry this proportion was estimated at 1.8%.[10] The missed cases in those studies had an almost zero chance of being notified to the cancer registry by pathology laboratories or hospitals. Therefore, the number of cases that were seen by the general practitioner only, or were treated abroad, could not be estimated by the capture-recapture method. By analogy, when zoologists use the capture-recapture method in a nature reserve, they do not pretend that they can estimate the number of animals living outside the fences of the nature reserve.

From the IKL study,[10] the proportion of missed cases that were omitted due to errors in the notification procedures could also be calculated. This proportion amounted to 2.2%.[10] This result was only slightly different from the estimate of the present study for the IKL cancer registry: 1.7%.

The capture-recapture method has been proposed by several authors for estimation of the completeness of prevalence studies or cancer registries.[2,4,5] The method is dependent on the assumption that the errors in notification sources are not correlated. If this assumption does not hold, it will influence the estimate of the unknown number of missed cases.

Whether or not the two notification sources in this study (PA and MR) were dependent, could not be derived from the analysis, because there were no known observations with $PA^{-}MR^{-}$. Even if a third notification source had been available, independence of the three notification sources cannot be proven either. The three-way interaction term in the modelling approach remains unknown.[11]

However, the pathology laboratories and medical departments are operated independently and the most likely explanation for missed registrations is, therefore, the existence of (uncorrelated) errors in the notification procedures.

If the sources are positively correlated, completeness will be overestimated, and vice versa. The correlation of errors can be expressed as the odds ratio (a*d/b*c).[4] When independence is assumed, the odds ratio is fixed at 1. In Table 4 the relation between the odds ratio and the estimate of completeness is shown.

TABLE 4 *Influence of correlated errors (dependence) in the notification sources (for several odds ratios) on the estimate of the completeness*

| Odds ratio | Total | |
|---|---|---|
| | Completeness % | 95% confidence interval |
| 0.5 | 99.1 | 98.9–99.4 |
| 1.0 | 98.3 | 97.9–98.7 |
| 1.5 | 97.4 | 96.9–98.0 |
| 2.0 | 96.6 | 95.9–97.3 |
| 2.5 | 95.8 | 94.9–96.7 |
| 3 0 | 95.0 | 94.0–96.0 |

In our opinion, the capture-recapture method is not a useful tool for monitoring the completeness of cancer registration on a routine basis. In contrast to the independent case ascertainment method, financial and personnel expenses are limited. However, the method will be useful only if it can be used for different sites, age groups, hospitals, etc. In general, deficiencies in notification procedures will occur in specific hospitals and perhaps also for specific sites. The number of strata in such an analysis will then be much larger than in the current study. Because of the limited numbers in the strata, the power of the method will become insufficient to detect deficiencies in notification procedures at an early stage.

It is therefore concluded, that the capture-recapture method is not useful for everyday surveillance of completeness.

REFERENCES

[1] Goldberg J, Gelfand H M, Levy P S. Registration evaluation methods: a review and a case study. *Epidemiol Rev* 1980; **2**: 210–20.

[2] Wittes J T, Colton T, Sidel V W. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chron Dis* 1974, **27**: 25–36.

[3] Neugebauer R. Application of a capture-recapture method (The Bernouilli census) to historical epidemiology. *Am J Epidemiol* 1984; **120**: 626–34.

[4] Hook E B, Regal R R. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol* 1992; **135**: 1060–67.

[5] Robles S C, Marrett L D, Clarke E A, Risch H A. An application of capture-recapture methods to the estimation of completeness of cancer registration. *J Clin Epidemiol* 1989; **41:** 495–501.

[6] Bishop Y M M, Fienberg S E, Holland P W. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press, 1975.

[7] Hook E B, Regal R R. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. *Am J Epidemiol* 1993; **139:** 1148–66.

[8] Baker R J, Nelder J A. *The GLIM System, Rev. 3,77.* Oxford: Numerical Algorithms Group, 1985.

[9] Berkel J. General practitioners and completeness of cancer registry. *J Epidemiol Community Health* 1990; **44:** 121–24.

[10] Schouten L J, Höppener P, van den Brandt PA, Knottnerus J A, Jager J J. Completeness of cancer registration in Limburg, the Netherlands. *Int J Epidemiol* 1993; **22:** 369–76.

[11] Kiemeney L A L M, Schouten L J, Straatman H. Ascertainment corrected rates (Letter to the editor). *Int J Epidemiol* 1994; **23:** 203–04.

*(Revised version received March 1994)*