Evaluation of the Completeness of Cancer Case Ascertainment in the Seoul Male Cohort Study: Application of the Capture-Recapture Method

Dae-Sung Kim¹, Moo-Song Lee², Dong-Hyun Kim³, Jong-Myon Bae⁴, Myung-Hee Shin⁵, Chung-Min Lee¹, Hye-Won Koo¹, Weechang Kang⁶, and Yoon-Ok Ahn¹

Since the completeness of case ascertainment is directly related to the validity of a study, the evaluation of completeness is an essential feature of a cohort study. To estimate the completeness of cancer case ascertainment during a three year period (Jan. 1, 1993, to Dec. 31, 1995) in which the Seoul Male Cohort was followed up, we applied capture-recapture method. Data were obtained from the cancer registries, medical records and death certificates, with cases identified from each source numbering 103, 105, and 38, respectively. After eliminating duplicate cases, the total number was 141, and by using a log-linear model, the number of cases not detected by any of the three data sources was estimated to be 16. For all cancers, the estimated completeness of follow-up was 89.9%. 

### MATERIALS AND METHODS

#### Study Population

To be eligible for inclusion in the study participants must be men resident in Seoul, insured of KMIC, aged from 40 to 59 at January 1991, and proven to be healthy in regular health examination which is conducted biennially as health promotion program of KMIC. The eligible population was 54,378 with 1,262 institutions. By cluster random sampling, 29,918
was sampled. We send postal questionnaires to 29,918 persons to obtain informations about lifestyles related to risk factors of cancer. Of the 29,189, 14,533 returned interpretable questionnaires (48.6%). The reason why we selected study subject from insured of KMIC is that each insured receives health examination biennially so that it is easy to construct disease-free cohort, and that the medical insurance claims of the insured is computerized and preserved for inspection so that we could easily get the data necessary for follow-up in cohort study.

Cancer case ascertainment from each data sources

Cancer cases were ascertained by three different ways(Figure 2). Medical insurance claim data was used for screening out potential cancer cases, about whom medical records was reviewed by visiting the hospitals. Data of cancer registries were obtained to detect cancer cases. Death certificate data was accessible from National Statistical office. Data linkage was done using resident identification number which is unique and given to all citizen of Korea. International classification of disease 140~208 (ICD-9) or C00~D09 (ICD-10) were regarded as relevant to cancer.

The medical insurance claims related to cancer were identified among cohort subjects. We were accessible to the computerized insurance claim data by courtesy of KMIC. The KMIC data consists of records of every hospital visit of each insured with their diagnosis as ICD cord, date of visit, whether inpatient or outpatient, cord number of visited hospital. We regarded subjects recorded to be cancer cases in claim data as potential cancer cases, not true cases, because the validity of medical insurance claim data is not proven yet. Therefore, we had to check the validity of medical claim data by visiting hospitals and reviewing medical records. The claim data relevant to cancer between January 1993 and December 1995 of claim data were reviewed. A total of 369 person were potential cancer cases; among these we accessed 301 medical records(Table 1). The medical records of 68 potential cases were inaccessible because of lack of cooperation from the hospitals or loss of medical records; of the 68 records, 10 were found at cancer registry data. remaining 58 potential cases were regarded as non-cases or missed cases. A structured abstract form was used for medical records review. Contents of the abstract form included first diagnosis of date, diagnostic measures and interpretation of the results, stage of cancer, treatment, etc. Well-trained reviewer(M.D.) decided whether the potential case had been a true cancer case or not by evaluating completed abstract form. The first visit recorded as cancer in KMIC data was regarded as an incidence date.

Cancer cases were also identified from two cancer registries, the Central Cancer Registry, which collects cancer cases from 121 nationwide training hospitals, and the Seoul Cancer Registry, which covers small hospitals in Seoul. Using these data, for the period of 1993 to 1995, cancer cases in the cohort were identified. The cancer registry data had not been merged with death certificate data. Thus, the cancer registry data used in this study did not include death certificate only(DCO) cases. In addition, mortality cases recorded in death certificates as cancer were identified. We used death certificate data from January 1993 through December 1996. To be used for capture-recapture method with other data sources of this study, incidence date should be known or otherwise estimated. We could not know the incidence date with death certificate data alone. Thus we estimated the incidence date using KMIC data assuming that first record in KMIC data as a cancer was incidence date. Fortunately, all mortality case were found to be potential cancer cases of KMIC data so that incidence dates of mortality cases were estimated using the first cancer claim date in the KMIC data.

Table 1. Number of persons who were recorded as a cancer case in medical insurance claim data of KMIC (1993-1995) and number of accessible medical records.

|        | 1993 | 1994 | 1995 | total |
|--------|------|------|------|-------|
| Potential cases | 124  | 122  | 123  | 369   |
| Reviewed       | 102  | 97   | 102  | 301   |
| Access rate    | 82.3%| 78.9%| 82.9%| 81.6% |
Completeness of Case Ascertainment

**Figure 1.** Distribution of cases by combination of three data source.

of two scaled deviances.

The goodness-of-fit of the three-source models were tested by likelihood ratio statistic. For log-linear model and parameter estimation, a GLIM 3.77 statistical package was used. In this study, G2 likelihood ratio goodness-of-fit was used to estimate the 95% confidence interval. To calculate goodness-of-fit based confidence interval, we used CATMOD procedure of SAS 6.12 which was more convenient for iteration than GLIM package.

To compare the estimates from various models, we applied two-source capture-recapture analysis. Each pairs from three data source was analyzed to estimate total population, for example, [medical record review] versus [cancer registries]. In addition, each combined pairs of data source and the other one was analyzed, for example, [medical records review and/or death certificates] versus [cancer registries]. Consequently, we tried six two-source models in addition to the three-source model.

If we had a expected value of cancer incidence, the validity of the estimate could be evaluated, though roughly. We calculated an expected incidence for three years of follow-up by indirect standardization using population-based incidence data of Seoul(1991). Cancer incidence data in Seoul beyond 1991 was not yet available. To take into account the change in age of subject, expected value was calculated in each years of follow-up with different age distribution and summated. The incidence data was from the implementation study of Seoul Cancer Registry, that didn't contain the DCO cases.

Finally, by comparing the detected number of cancer cases with the estimated total cases, we calculated the completeness of ascertainment from medical records, cancer registry data, death certificates and overall case ascertainment.

\[
\text{observed completeness} = \frac{\text{detected cases in each data source}}{\text{observed cancer cases}}
\]

\[
\text{estimated completeness} = \frac{\text{detected cases in each data source}}{\text{estimated cancer cases}}
\]

**RESULTS**

**Number of cancer cases ascertained by each data sources**

From a review of the 301 medical records by visiting hospitals, 105 incident cancer cases were identified to be occurred between 1993 and 1995. From cancer registry data, 103 cases were ascertained during the same period. From death certificate data, 38 cancer cases were identified. There was gradual increase in proportion of cases identified from cancer registries whereas the number of cases from other data sources showed somewhat fluctuating pattern.

Other than duplicates, a total of 141 cancer cases were found (Figure 2, Table 2). There was no remarkable increase of annual number of cancer incidence. Medical records review ascertained 74.5% of total cases. Number of cases found in cancer registries was not much different from that ascertained by medical records review. However, only 27.0% of total cases were found in death certificate data.

All cases, as identified by combining the three data sources, are shown in Table 3. Sixteen cases (11.3%) were detected from all three sources; 73 (51.8%) were detected from two, and 52 (36.9%) from one only. When combine each couple of sources, medical records review and cancer registry detected 140 cases which is 99.3% of total cases. In combination of medical records review with death certificates and cancer registries with death certificates, 123 and 109 cases were detected respectively.
**Table 2.** Number of cases in Seoul Male Cohort identified by three data sources.

| data sources         | number of cancer cases |
|----------------------|------------------------|
|                      | 1993  | 1994  | 1995  | total |
| Medical record       | 37 (74.0%) | 31 (70.5%) | 37 (78.7%) | 105 (74.5%) |
| Ascertainment        |       |       |       |       |
| Cancer registry      | 33 (66.0%) | 31 (70.5%) | 39 (83.0%) | 103 (73.0%) |
| Death certificate    | 12 (24.0%) | 17 (38.6%) | 9 (19.1%)  | 38 (27.0%)  |
| Combined             | 50 (100%) | 44 (100%) | 47 (100%)  | 141 (100%)  |

**Figure 2.** Flow of cancer case ascertainment from 1993 through 1995.

*Model fitting and Estimation of total cancer cases by capture-recapture method*

Table 4 is result of test about whether heterogeneity is present or not. Expected cases were calculated from exponential of parameter from two model; 'A' is a result of three-source analysis using all data; 'B' is a result of three-source analysis using data excluding cases who are detected from all of three sources of data. The expected numbers are similar in both A and B model, except in the number of case detected from all of three sources of data; the expected number from model B is 3.0 more than that from model A. However, the difference of deviance between two model was not significant.
The result in Table 5 shows the statistical significance of each data source in three-source log-linear models. After adding the interaction term, there was significant decrease of scaled deviance. The significant interaction terms were 'medical record review + death certificates' and 'cancer registries + death certificates'. However, when the interaction term 'medical record review + death certificates' was included in the model, there was a loss of significance of the main effect of 'medical record review'. Finally, the model including the interaction of 'cancer registries + death certificates' was found to be most fitted. It means that there was dependence between cancer registry data and death certificate data. The number of undetected cases was estimated by summation of exponential function of parameters estimated. According to the model selected in Table 5, the total number of cancer cases were estimated to be 156.8, which is 15.8 more than the total detected cases from the three data sources and 4.2 more than total cases estimated by three-source model without interaction term.

We presented various estimates using two-source models and three-source models (Table 6). Two-source models cannot solve the problem of dependence. If so, in two-source models with positively dependent sources, there would be underestimation. If the positive dependence shown in three source analysis is true, model 3, 5 and 6 (Table 6) is two-source model with positive dependence unsolved, and the estimated total cases would be underestimated. Actually, the estimated total cases from model 3, 5 and 6 was fewer than that of both three-source models with and without interaction term. The minimal estimate was from model 3 which consists of two data sources dependent on each other.

To evaluate the estimated cancer incidence, standardized incidence number of cases was calculated from cancer incidence data of Seoul, 1991 (Figure 3). The expected value (155.9) was near the estimated value and far from observed value. Comparing with expected value, the estimation from three-source model with interaction was found to be more reasonable than that from the other models.

Completeness of case ascertainment

Table 7 illustrates the completeness of each data source and
Table 5. Log-linear model fitting and evaluation for parameters and dependencies in three data sources-all cancers.

| Model* | Scaled deviance | d.f. | significant variables** |
|--------|-----------------|------|-------------------------|
| 1+2+3  | 13.710          | 3    | 1, 2, 3                 |
| 1+2+3+(1*2) | 9.928          | 2    | 1, 2, 3                 |
| 1+2+3+(1*3) | 9.047          | 2    | 2, 3, 1*3               |
| 1+2+3+(2*3) | 6.043          | 2    | 1, 2, 3, 2*3            |
| 1+2+3+(1*2)+(1*3) | 1.577        | 1    | 2, 3, 1*2, 1*3          |
| 1+2+3+(1*2)+(2*3) | 5.183          | 1    | 1, 2, 3, 2*3            |
| 1+2+3+(2*3)+(1*3) | 2.489          | 1    | 2, 3, 2*3               |

*1-medical record ascertainment, 2-cancer registry, 3-death certificate data
**p-value < 0.05

Table 6. Comparison of various models and total cancer occurrence estimated.

| models                  | Number of cancer cases | (95% C.I.)  |
|-------------------------|------------------------|-------------|
|                         | Observed  | estimated  |             |
| two-source analysis     |           |            |             |
| 1. MR vs. CR            | 140       | 160.1      | (149.4-176.2)|
| 2. MR vs. DC            | 123       | 196.9      | (151.7-242.0)|
| 3. CR vs. DC            | 109       | 126.8      | (113.4-140.1)|
| 4. MR vs. (CR or DC)    | 141       | 159.7      | (148.5-170.9)|
| 5. CR vs. (MR or CR)    | 141       | 152.3      | (144.7-159.8)|
| 6. DC vs. (MR or DC)    | 141       | 148.6      | (142.4-154.8)|
| three-source analysis   |           |            |             |
| 7. MR vs. DC vs. CR     | 141       | 152.6      | (146.5-163.3)|
| 8. MR vs. DC vs. CR*    | 141       | 156.8      | (147.6-170.7)|

abbreviation: MR, medical records ascertainment; CR, cancer registries; DC, death certificates
* model with interaction term of DC*CR

Figure 3. Observed incidence, estimated incidence from log-linear model and expected incidence from indirect standardization using populational cancer incidence data of Seoul, 1991.
Completeness of Case Ascertainment

Table 7. Completeness of data sources by three-source capture-recapture estimate with interaction term.

| Data source              | Completeness (%) | (95% C.I.)   |
|--------------------------|------------------|--------------|
|                          | Observed         | estimated    |              |
| Medical records review   | 74.5             | 67.0         | (61.5-71.1)  |
| Cancer registries        | 73.0             | 65.7         | (60.3-69.8)  |
| Death certificates       | 27.0             | 24.2         | (22.3-25.7)  |
| Overall case ascertainment| 88.1             | 89.9         | (82.6-95.5)  |

Overall case ascertainment. Observed completeness used 141 detected total cases and estimated completeness used 156.8 estimated from three-source model with interaction term. Estimated completeness of medical records review and cancer registries was over 60%. But the completeness of death certificate data was below 25%. Estimated completeness of overall case ascertainment was almost 90%. The confidence interval of completeness was calculated from that of estimated total cancer cases.

DISCUSSION

Among the total of 369 potential cancer cases from KMIC data, 301(81.6%) medical records were accessible. However, only 105 were ascertained to be a true cancer by medical record review. This high miscoding rate (65.1%) would be likely due to following reason. In Korea, medical claim of expensive and redundant diagnostic measures need more persuasive, even though false, ICD code rather than that of real diagnosis. Nowadays, several validity studies are being conducted to prove this phenomenon, but not published yet. However, 196 miscoded subjects does not influence the result because they were not included in capture-recapture analysis. Although there may be true cases among the miscoded, they would not lead biases since the capture-recapture method itself was based on the presence of missed case and tried to find out the size of them.

Validity of data is important in capture-recapture method, in other words, there should be no false-positive cases. Misdiagnosis can over- or underestimate the result of the capture-recapture method. In particular, if there is a false positive case, correction for underestimation is not perfect. The cases ascertainment by medical records review have few problem of false-positive case, because we judged the validity by various aspect of each record, for example, diagnostic tools with interpretation, treatment done, stage of cancer, etc.

Cancer registry data used in this study didn't include DCO. All cases in the cancer registries were reported from hospitals with histological diagnosis, first date of diagnosis and pathologic diagnosis, etc. The proportion of histological verification(HV%) was 77% in Seoul Cancer Registry(1991), which were slightly lower than other reliable cancer registry(94% for whites in US Connecticut, 95% for whites in US SEER, 70 and 73% for male and females in Osaka, Japan). The lower levels of HV% might be due to the high relative frequency of liver cancer in Korea, In fact, the HV% of liver was 26% in males, while that of other major cancer was over 85%.

In case of mortality data, there could be a problem of validity. In nationwide death certificate data in 1995, cause of death was unclassifiable in 0.8%, medical doctor certified the cause of death only in 60.6% 13). However, in this study, all the 38 cases from death certificate data was certified by medical doctor. The high certification rate may be likely due to relatively young age of subjects and the attribute of study subjects who were all public servents and residents of metropolitan.

The reason for relatively small number of cases detected from death certificate is that we used the death certificate data of only 4 years, from 1993 through 1996. When we can get the data after 1996, the cases of first three years of follow-up will be more than 38.

The capture-recapture methods involves several assumptions 5). First, the population should not change during the investigation; secondly, there should be no loss of tags, and thirdly, for each sample, each individual should have the same chance of being included; fourthly, the two samples should be independent.

When the capture-recapture method is applied to a cohort study, the first assumption is usually true because there is no temporal order between capture and recapture in case ascertainment by multiple sources 14). Migration out of jurisdiction, death, and other loss of follow-up can also violate first assumption in fixed cohort study 14). However, the three sources of data used in this study are nationwide databases so that migration out of Seoul doesn't mean loss from follow-up. Death cannot be a loss of follow-up because we used death certificate data. One reason of loss from follow-up may be emigration, but it could not be substantial in cases or potential cases enough to be considered in this study.

The second assumption is completely satisfied in Korea, since every citizen has their own identification number, and it was included in all the three sources of data without missing.

The third and fourth assumptions are more often violated.
The third is that every individual should have the same chance of being included in the sample if not, there is a heterogeneity which implies that dependency is present in all possible combinations of data sources. The failure of third assumption is due to different probability of being caught in any sample. In this study, if some of the cohort subjects were less more probable of being caught by any of three data sources, there might be a heterogeneity. Heterogeneity depends on the character of individual not of sources of data. The heterogeneity can be test in various way, but we tried a brief test where we found no evidence of heterogeneity. Heterogeneity can also be considered with regard to biological consideration. We could not suppose that any group among the cancer cases of this study had different probability of being caught from the other group, because the study subjects have same race, same sex, narrow range of age and not so much different socio-economic status.

The fourth is that different case, respectively detected and not detected in the same data sources, have the same probability of being detected in another data source. For example, if cancer cases detected by cancer registry are more likely to be detected by a review of medical records than not detected cases in the cancer registry, the fourth assumption is violated. In this instance, two sources of data show positive dependency, and this underestimates the number of undetected cases. Conversely, negative dependency may occur, and this overestimates the number of undetected cases. If there are three sources of data, two-source dependence (first order interaction) can be solved by fitting model with interaction term. The dependence between cancer registries and death certificates was testified by fitting log-linear models, although second order interaction could not be testified. Another method to test dependence, as Wittes et al. suggested, is that, if there are more than two sources, one can compare the sources two at a time. If one estimates is much lower than the rest, then one might suspect positive dependency. In this study, we performed two-source analysis to support the presence of dependence. Two-source analysis of Cancer registry with death certificate data show the less estimate than any other two-source analysis, which supports the presence of positive dependency as appeared in three-source log-linear analysis. When two sources of data combined to one sample and twosource capture-recapture analysis was conducted with another, two models including dependent sources in each side showed less estimate than that of another two-source analysis and three-source analysis with interaction term. Thus, we believe that positive dependence between cancer registries and death certificate data is appropriately treated by log-linear model.

Since we had no true value to be compared with, the population-based cancer incidence rate of Seoul was projected into this study population to get a expected value. The estimated value was almost as expected value. When considering that the population-based cancer registry is Seoul Cancer Registry which we used in case ascertainment, the number of cases, 103, from cancer registries is much lower than the expected value, 155.9, from the Seoul Cancer Registry. The much lower incidence in study population can be explained that study subjects was healthy population at the time of beginning of follow-up so that prevalence of risk factors for cancer might be lower than that of general population. With regard to completeness, if the estimated completeness of cancer registries could be generalized, the completeness-adjusted expected incidence cases in this cohort would be 237.3 which is 155.9(expected cases) over 0.657(completeness of cancer registries). Thus, the interpretation of Figure 3 should be limited.

Exhaustive case ascertainment can definitely not be replaced by the capture-recapture method, but without the method, it is very difficult to evaluate the completeness and effectiveness of follow-up in a cohort study. This study showed that the capture-recapture method can be used to evaluate underestimation of the incidence rate calculated during such a study.

REFERENCES

1. LaPorte RE, Dearwater SR, Chang YF, et al. Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring. Am J Epidemiol, 1995; 142: 1069-1077.
2. Hook EB, Regal RR. 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: 1060-1067.
3. McCarty DJ, Tull ES, Moy CS et al. Ascertainment corrected rates: applications of capture-recapture methods. Int J Epidemiol, 1993; 22: 559-565.
4. Yoon-Ok Ahn. Diet and stomach cancer in Korea. Int J Cancer, 1997; 10: 7-9.
5. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. Am J Epidemiol, 1995; 142(10): 1047-1058.
6. The GLIM Working Party. The Generalized Linear Interactive Modelling system, Release 3.77. Oxford, Numerical Algorithm Group, Ltd., 1987.
7. Cormack RM. Log-linear models for capture-recapture. Biometrics, 1989; 45: 395-413.
8. Ronald RR, Ernest BH. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. Stat Med, 1984; 3: 287-291.
9. Cormack RM. Interval estimation for Mark-Recapture
studies of closed populations. Biometrics, 1992; 48: 567-576.
10. SAS/STAT User's guide, release 6.03. SAS Institute Inc, 1991.
11. Shin MH, Ahn YO. Evaluation of the completeness and validity of the registration in the Implementation Study of Seoul Cancer Registry (ISSCR). Kor J Prev Med, 1994; 27(4): 735-745.
12. Brenner H. Effects of misdiagnoses on disease monitoring with capture-recapture methods. J Clin Epidemiol, 1996; 49: 1303-1307.
13. National Statistical Office. Annual report on the cause of death statistics, 1995.
14. Hook EB, Regal RR. Internal validity analysis: A method for adjusting capture-recapture estimates of prevalence. Am J Epidemiol, 1995; 142(S9): s48-s52.
15. Darroch JN, Fienberg SE, Glonek GF, Junker BW. A three-sample multiple-recapture approach to census population estimation with heterogeneous atchability. J Am Stat Assoc, 1993; 88(423): 1137-1148.
16. Wittes JT, Colton T, Sidel VW. Capture-recapture models for assessing the completeness of case ascertainment using multiple information sources. J Chronic Dis 1974; 27: 25-36.