ORIGINAL ARTICLE
Oncology & Hematology

JKMS

http://dx.doi.org/10.3346/jkms.2012.27.5.489 • J Korean Med Sci 2012; 27: 489-494

Completeness of Cancer Case Ascertainment in Korea Radiation Effect and Epidemiology Cohort Study

Minkyo Song[1], In-Seong Cho[1], Zhong Min Li[2], and Yoon-Ok Ahn[1,2]

[1]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul; [2]Institute of Radiation Effect & Epidemiology, Seoul National University Medical Research Centre, Seoul, Korea

Received: 26 July 2011
Accepted: 23 February 2012

Address for Correspondence:
Yoon-Ok Ahn, MD
Department of Preventive Medicine, Seoul National University College of Medicine, 103 Daehank-ro, Jongno-gu, Seoul 110-799, Korea
Tel: +82.2-740-8322, Fax: +82.2-747-4830
E-mail: yoahn@plaza.snu.ac.kr

The aim of this study was to evaluate whether the completeness of case ascertainment during the follow-up of a cohort differed between the exposed and the nonexposed groups in Korea Radiation Effect and Epidemiology Cohort (KREEC). The completeness was defined as the proportion of the number of detected cases to the number of estimated cases, in which the estimation was performed by capture-recapture method. Data were obtained from the cancer registries, death certificates, and medical records during years 2004-2007. Among 11,367 subjects in the exposed group and 24,809 subjects in the unexposed group, the completeness of cancer case ascertainment were 88.2% vs 87.2% in cancer registry, 38.2% vs 41.1% in death certificate and 57.9% vs 62.0% in medical records data, 96.9% vs 97.1% for all combined sources and were not statistically different between the two groups. In conclusion, the method of ascertaining the cases in the KREEC was not biased depending on the exposure status, and thus adds credibility to the outcomes of the KREEC study as well as confirming the incident cases in the two groups.

Key Words: Cohort Studies; Data Collection; Neoplasms; Validity

INTRODUCTION

In a cohort study, it is essential to balance the following up of the selected exposed group and nonexposed group. The major source of bias that occurs during the measurement of outcome may occur from inadequate means of obtaining information regarding exposures and/or disease outcome. In other words, if there is a difference in the completeness of case-ascertainment between the two groups depending on the exposure, for instance, a surveillance bias in which disease ascertainment may be better in the monitored group than in the general population- an ascertainment bias-, the incidence or other outcome variables that result from the study-the relative risk or odds ratio-become incredible (1, 2).

Various data sources are used in detecting the occurrence of target disease. They can be categorized into either active or passive follow-ups. Passive follow-up is to achieve data collected and maintained by organizations outside the study for other purposes, and the examples would be cancer registry, death certificate data, health insurance data, etc. Active follow-up requires direct contact with the cohort participants by mailings, phone calls, interviews, etc. Collecting certain types of information using standardized case report forms from hospital records that were not originally purposed for ascertaining outcomes could also be considered as active follow-ups. However, no one data source or combinations of multiple data sources render a full detection of the cases, due to the cases that are missed in each data sources. Thus, various methods have been developed to supplement in estimating the accurate incidence of disease, one being capture-recapture method. Originally developed for counting fisheries and wildlife animals, capture-recapture method has later been employed in many epidemiologic researches to estimate the unobserved cases and evaluate the completeness of data from various incomplete data sources. Yet, application of the capture-recapture method requires several assumptions that can be elusive in real life epidemiologic studies that may be overcome by the use some mathematical models, and an example would be the log-linear model (3).

The purpose of this study was first, to estimate the completeness of cancer cases in Korea Radiation Effect and Epidemiology Cohort (KREEC) during the follow-up period of 2004 to 2007, using cancer registry data, death certificate data and medical records obtained actively from hospitals, and secondly to evaluate whether there is a difference between the completeness of case ascertainment between the exposed group and the nonexposed group during the period.

MATERIALS AND METHODS

Study population

KREEC study has been initiated in 1992 to scientifically evaluate the health effects of radiation emitted from the nuclear power plants in Yeong Gwang, Korea, on the residents who res ide near the plants. Exposed group was composed of workers at the

pISSN 1011-8934
eISSN 1598-6357

power plants and residents living within 5 km radius, and the nonexposed group was set at two different levels - intermediate-distance group of residents living 5-30 km away from the power plants, and far-distance group of residents living more than 30 km away from the power plants. Upon selecting members based on strict criteria, the follow-ups had been conducted by three different methods since 1993 (Fig. 1). Cancer registry data, death certificate data and medical records from the participating hospitals were used to detect cancer cases defined as C00-D09 of International Classification of Diseases-10 (ICD-10). Active check-ups of medical records were performed in those who have been identified suspicious of having cancer - all cases having C00-D09 codes, and randomly selected cases with other codes from the National Medical Claims Data. Medical doctors have reviewed the structured abstracts to evaluate the cancer cases, constructed from medical records by trained interviewers, including diagnosis of date, diagnostic measures and interpretation of the results, stage of cancer, treatment, etc.

Case definition

A cancer case was defined as having a cancer diagnosis through biopsy, cytology, CT/MRI, and other imaging procedures in medical records data. The date of cancer diagnosis was regarded as the incident date. Cancer code and date were also identified either from Korea Central Cancer Registry (KCCR) or death certificate data, whichever the cancer diagnosis date was earlier. The analysis of the study was conducted using data from year 2004-2007, since the cancer cases before year 2004 is thought to be complete for both the exposed and nonexposed groups, and some of the cases might include prevalent cancer cases. Finally, the subjects in included were 11,367 near residents (supposedly exposed group) and 24,809 intermediate and far residents (supposedly nonexposed group).

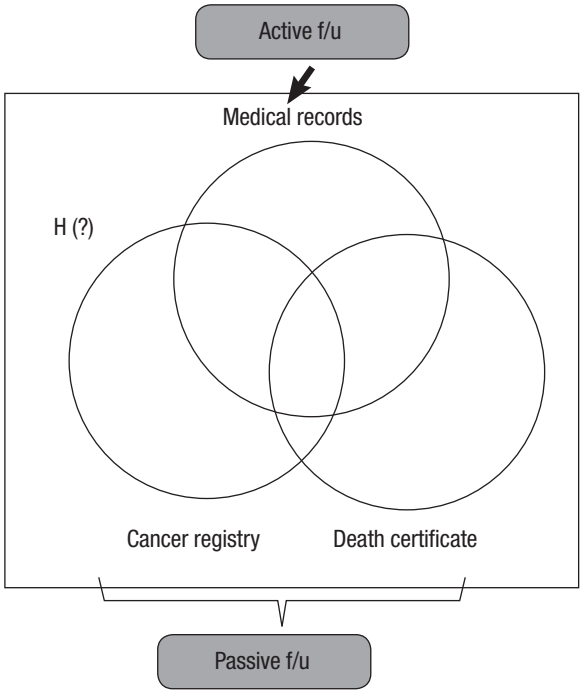

Fig. 1. Data sources used for case-ascertainment during follow-up (f/u) of Korea Radiation Effect and Epidemiology Cohort Study (KREEC). H (?) represents cases not detected in any of the three data sources.

Statistical analysis

To compare the estimated number of cancer cases in different models, two-source capture-recapture analysis was conducted, in which cancer registry data vs death certificate data, death certificate data vs medical records data, and medical records data vs cancer registry data are used. Assuming the dependence of each data sources, maximum likelihood estimator (MLE) was used for each of the comparing two sources to estimate the complete number of cases and its confidence interval, and indirectly evaluate the dependency between the sources (4).

The unobserved 'H' in Fig. 1 can be estimated using log-linear models in three-source model. In order to use the log-linear model, several assumptions should be met, including the independence between the data sources. To resolve the independence, interaction terms were added in the models for examining and adjusting dependency (3).

Finally, estimated number of cases was compared to the observed cases to evaluate the completeness of case-ascertainment during the follow-up in KREEC study, as the completeness is defined as the proportion of the number of cases detected to the number of estimated cases during the follow-up.

RESULTS

Cancer cases ascertained by each data sources

Table 1 shows detected cancer cases in three different data sources. A total of 332 cancer cases were detected in cancer registry data, 144 cases in death certificate data, and 218 cases were confirmed through medical records in near residents. In intermediate/far residents, 755 cases, 356 cases and 537 cases were detected respectively for each data sources. Considering the cases that were found in multiple sources, a total of 365 cases of 11,367 near residents (3.2%), and a total of 841 cases out of 24,809 intermediate/far residents (3.4%) were identified through cancer registry, death certificate and medical records data.

Table 1. Cancer cases detected in three data sources in Korea Radiation Effect and Epidemiology Cohort (KREEC) during follow-up 2004-2007

Data sources	Near	Intermediate/Far
	No. (%)	No. (%)
Cancer registry	332 (91.0)	755 (89.8)
Death certificate	144 (39.5)	356 (42.3)
Medical records	218 (59.7)	537 (63.9)
Total	365 (100.0)	841 (100.0)

http://dx.doi.org/10.3346/jkms.2012.27.5.489

Fig. 2 shows more specified number of cases detected in three different data sources for each area. The numbers of cases are missing for values that are detected in none of the data sources for both near residents and intermediate/far residents.

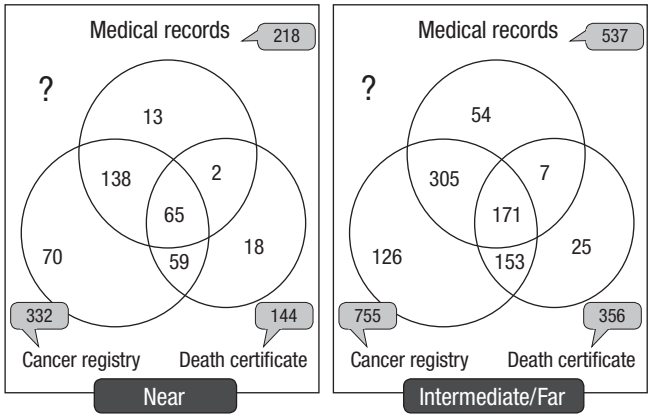

Fig. 2. Detected cancer cases in near residents and intermediate/far residents during follow-up of 2004-2007.

Comparison of number of cancer cases from each data sources

The incident rates identified in each data sources are shown in Table 2. Chi-square test for heterogeneity among three sources was not significant ($P > 0.05$) between near residents and intermediate/far residents demonstrating no statistically significant difference between exposed group and the nonexposed group in the cohort.

Estimation of cancer cases using two-source method

Table 3 is a result of calculated estimation of cancer cases from two-source data sources. Results from data that are matched to cancer registry (cancer registry vs death certificate and medical record vs cancer registry) rendered little difference between the observed number and the estimated number of cancer cases, whereas values from death certificate vs medical records data produced the estimated numbers of cases ranging from 1.0 to 1.6 times the number of observed cases. The data from the three different combinations of data sources were not significantly different between near residents and intermediate/far residents.

Table 2. Observed incident rates in near residents and intermediate/far residents in three data sources during 2004-2007

Area	No. of residents	Cancer registry		Death certificate		Medical records	
		No.	IR*	No.	IR*	No.	IR*
Near	11,367	332	29.2	144	12.7	218	19.2
Int/far	24,809	755	30.4	356	14.3	537	21.6
Total	36,176	1,087	30.0	500	13.8	755	20.9

*IR = incidence rate (n/1,000 persons). χ^2 test for heterogeneity among three sources was not significant ($P > 0.05$). int/far, intermediate/far.

Table 3. Estimation of cancer cases in two-source (2 × 2 comparison)

Area	CR vs DC		DC vs MR		MR vs CR	
	Observed No.	Estimated No. (95% CI)	Observed No.	Estimated No. (95% CI)	Observed No.	Estimated No. (95% CI)
Near	352	386 (366-406)	295	469 (400-537)	347	357 (349-365)
Int/Far	787	830 (809-850)	715	1,074 (983-1,165)	816	852 (836-867)
Total	1,139	1,213 (1,181-1,241)	1,010	1,541 (1,428-1,654)	1,163	1,209 (1,191-1,226)

95% CI based on asymptotic normal distribution. int/far, intermediate/far; CR, Cancer Registry; DC, Death Certificate; MR, Medical Records.

Table 4. Log-linear model fitting and evaluation for parameters and dependencies in three data sources in study population

Area	Model*	Scaled deviance	d.f.	Significant variables†
Near residents	1+2+3	29.2493	3	1 ,2, 3
	1+2+3+(1*2)	10.4581	2	1, 2, 3, 1*2
	1+2+3+(1*3)	28.0572	2	1, 2, 3
	1+2+3+(2*3)	16.3965	2	1, 3, 2*3
	1+2+3+(1*2)+(1*3)	6.3018	1	1, 2, 3, 1*2, 1*3
	1+2+3+(1*2)+(2*3)	2.6430	1	1, 3, 1*2, 2*3
	1+2+3+(1*2)+(2*3)+(1*3)	14.3180	1	1, 3, 2*3
Intermediate/far residents	1+2+3	57.0555	3	1 ,2, 3
	1+2+3+(1*2)	52.7060	2	1, 2, 3, 1*2
	1+2+3+(1*3)	47.3110	2	1, 2, 3, 1*3
	1+2+3+(2*3)	17.9938	2	1, 3, 2*3
	1+2+3+(1*2)+(1*3)	25.6873	1	1, 2, 3, 1*2, 1*3
	1+2+3+(1*2)+(2*3)	17.1093	1	1, 3, 2*3

*1, cancer registry; 2, medical records; 3, death certificate; d.f., degree of freedom. †$P < 0.05$.

Model fitting using log-linear models

Table 4 illustrates the results of statistical significance of each data sources in three-source log-linear models. Adding interaction term decreased the scaled deviance. However the final model that was found to be most fitted was the one that included no interaction terms. This model demonstrates that there was no dependence between any of the data sources. According to the model selected in Table 4, the estimated numbers of cancer cases were 376.6 in near residents and 865.8 in intermediate/far residents as shown in Table 5.

Completeness of case ascertainment in exposed and nonexposed groups

Finally the completeness of case-ascertainment was considered between the exposed group (near residents) and the nonexposed group (intermediate/far residents). The completeness of case-ascertainment was calculated as the percentage of observed number of cases out of estimated number of cases in each group in Table 6. The completeness of cancer registry data was approximately 88% in cancer registry data, 60% in medical records data, and 40% in death certificate data. However, the completeness of each data sources between the two groups had shown no statistical difference ($P = 0.72$) in a chi-square test. Likewise, the completeness calculated in all three sources combined rendered 96.9% in exposed group and 97.1% in nonexposed group with no statistically significant difference ($P = 0.84$). In conclusion, there was no statistical difference in the completeness of cancer case-ascertainment between the exposed and the nonexposed groups during the follow-up of the cohort study.

DISCUSSION

Completeness of incidence in cancer registries can be achieved using various available methods, such as death certificate cases (DNC) method, mortality/incidence ratio, historical comparison, Bullard method, etc (5, 6). In this study, we have used capture-recapture method to estimate the completeness of cancer case ascertainment and to finally evaluate whether the completeness differed between exposed and nonexposed groups in the Korea Radiation Effect and Epidemiology Cohort, and showed that the completeness of the three data sources were not statistically different between near resident (exposed group) and intermediate/far resident (nonexposed group) (Table 6). Despite some of its limitations, capture-recapture method in estimating the completeness in diseases has been widely used in the field of epidemiology (3, 7-10). The advantages and disadvantages in using capture-recapture method to estimate cancer cases in cohorts, using three sources-cancer registry, death certificate, and medical records in Korea-is well explained in a study performed in the Seoul Male Cohort Study (11).

Nonetheless, whatever the method, the estimated completeness do not reach 100%, as seen in the results of this study, but if the follow-up period is long enough the completeness should eventually approximate 100%. Thus, the data before year 2004 in the cohort were excluded, since case-ascertainment during the period was thought to be complete and was no longer subject to evaluation. Subjects who were suspicious of having cancer - those having C or D codes in medical claims data - and some randomly selected individuals with other codes during 2004 and 2007 were reviewed for medical records. In reality, not all members of a cohort study can be actively followed-up, which is the most accurate method in validating a case, due to immense cost and effort required, most cohort studies adopt passive surveillance system to supplement the completeness of follow-up (12). Likewise, this study has also used cancer registry and death certificate data to identify those who could have been missed as cases due to various reasons.

Consequently, completeness of case-ascertainment is critical in the evaluation of the results and the completeness must be guaranteed to be identical in both exposed and nonexposed groups. Hence, before analyzing relative risks or odds ratios, it is mandatory to first consider the validity of completeness in the ascertainment of cases in the following-up of a cohort study. Several studies in Korea have evaluated the completeness of case ascertainment using capture-recapture method, but they were limited to the analysis of completeness of the entire study

Table 5. Comparison of various models and cancer occurrence estimated in near residents and intermediate/far residents

Area	Models	Obs. No.	Est. No.	(95% CI)
Near	CR vs MR vs DC	365	376.6	(372.8-382.2)
Int/Far	CR vs MR vs DC	841	865.8	(860.3-872.9)
Total	CR vs MR vs DC	1,236	1,242.4	(1,235.5-1,251.1)

Int/Far, intermediate/far; CR, Cancer Registry; DC, Death Certificate; MR, Medical Records; Obs., observed; Est., estimated.

Table 6. Completeness of data sources by three-source capture-recapture method in exposed group vs nonexposed group

Area	Est. No.	CR†			DC†			MR†			All†		
		Obs. No.	Com.*	(95% CI)	Obs. No.	Com.*	(95% CI)	Obs. No.	Com.*	(95% CI)	Obs. No.	Com.*	(95% CI)
Near	376.6	332	88.2	(86.9-89.1)	144	38.2	(37.7-38.6)	218	57.9	(57.0-5.5)	365	96.9	(95.5-97.9)
Int/Far	865.8	755	87.2	(86.5-87.8)	356	41.1	(40.8-41.4)	537	62.0	(61.5-62.4)	841	97.0	(96.3-97.8)
Total	1,242.40	1,087	87.5	(86.9-88.0)	500	40.2	(40.0-40.5)	755	60.8	(60.3-61.1)	1,236	99.5	(98.8-100.0)

*Completeness = {(observed number of cases)/(estimated number of cases)} × 100. †P value calculated by chi-square test between the near and int/far groups; CR/DC/MR ($P = 0.72$), all ($P = 0.84$). int/far, intermediate/far; CR, Cancer Registry; DC, Death Certificate; MR, Medical Records; Obs., observed; Est., estimated; Com., Completeness.

http://dx.doi.org/10.3346/jkms.2012.27.5.489

population (11, 13-15). Likewise, in various cohort studies in Korea, which have used nested case-control study for analysis, researchers have disregarded or have omitted to mention to evaluate the validity of completeness of case-ascertainment in exposed and nonexposed groups before selecting case and control groups, therefore prone to selection bias (7, 16, 17).

In this study the completeness of each data source has ranged from as low as approximately 38.2% to as high as 88.2% (Table 6). Completeness of cancer registry data was highest in both near residents and intermediate/far residents. Compared to the previous study using three-source capture-recapture method performed with the registry data of 1993-1995, the completeness of cancer registry is higher in this study (87%-88% vs 67%), perhaps due to the actual improvement of detection and/or reporting rate (11). Low completeness of 38.2%-41.1% estimated with death certificate may owe to the use of only 4 yr of the data, from 2004 through 2007. The completeness of three-source method-applied estimate of 99.5% is similar to the number of 94.6%, the estimated number of nationwide cancer incidence by the Ajiki method calculated for the years 2003-2005 (18).

The missed cases, i.e. the "H" from Fig. 1 are speculated to be cancer cases that are not detected by the three sources used in the study due to several reasons. One reason being that these cases are actually patients who are at "before diagnosis status" because they have not visited the hospital yet, or have died due to reasons other than cancer. Another possible scenario would be cancer cases that are diagnosed elsewhere, hospitals that are not registered for the Central Cancer Registry system, or at hospitals abroad.

The result of this study has suggested that there was no statistically significant difference between the exposed and nonexposed groups in the ascertainment of cancer. An important consideration in using the capture-recapture method in epidemiologic data is that most often data sources are not independent. These positive dependencies which may underestimate true number of missed cases were compensated through log-linear models with three data sources. Furthermore, the three data sources are nationwide databases based on the resident registration number, so that migration out of the area does not mean loss from follow-up. Nonetheless, although the numbers for individual group (exposed and nonexposed) were thought to be estimated without bias, careful interpretation is needed in comparing the two estimated values. That is even if there is no heterogeneity between the two groups, there still remains a chance that within the groups the probability of being caught might differ. With regard to biologic considerations, exposed or nonexposed groups can over-report or under-report their conditions. However, since the purpose of this study aimed at examining the catchability between the two groups, this is not much of a consideration.

In evaluating the association between the exposure and the outcome in a prospective study, the fundamental assumption to be qualified is that there should not be a difference in ascertaining the cases between the exposure groups. The result of our study showed that the completeness of data sources derived from observed number of cases over estimated number of cases by capture-recapture method were 96.9% in the exposed group (near residents) and 97.1% in the nonexposed group (intermediate/far residents), which were not significantly different. This result also adds credibility to the outcomes of the KREEC study, as well as confirming the incident cases in the two groups.

REFERENCES

1. Gordis L. *More on causal inferences: bias, confounding, and interaction. In: Epidemiology. 4th ed. Saunders, Elsevier, 2009, p 247-63.*
2. Ahn YO, Yoo KY, Park BJ, Kim DH, Bae JM, Kang D, Shin MH, Lee MS. *Errors and bias in research results. In: Ahn YO, editor. Epidemiology: the principles and applications. Seoul: Seoul National University Press, 2005, p 307-31.*
3. 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: 1047-58.*
4. Regal RR, Hook EB. *Goodness-of-fit based confidence intervals for estimates of the size of a closed population. Stat Med 1984; 3: 287-91.*
5. Schmidtmann I, Blettner M. *How do cancer registries in Europe estimate completeness of registration? Methods Inf Med 2009; 48: 267-71.*
6. Tilling K. *Capture-recapture methods: useful or misleading? Int J Epidemiol 2001; 30: 12-4.*
7. International Working Group for Disease Monitoring and Forecasting. *Capture-recapture and multiple-record systems estimation II: applications in human diseases. Am J Epidemiol 1995; 142: 1059-68.*
8. Hook EB, Regal RR. *Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995; 17: 243-64.*
9. Neugebauer R, Wittes J. *Voluntary and involuntary capture-recapture samples: problems in the estimation of hidden and elusive populations. Am J Public Health 1994; 84: 1068-9.*
10. Brenner H, Stegmaier C, Ziegler H. *Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany. J Epidemiol Community Health 1995; 49: 426-30.*
11. Kim DS, Lee MS, Kim DH, Bae JM, Shin MH, Lee CM, Koo HW, Kang W, Ahn YO. *Evaluation of the completeness of cancer case ascertainment in the Seoul male cohort study: application of the capture-recapture method. J Epidemiol 1999; 9: 146-54.*
12. Lee MS, Kang WC, Kim DH, Bae JM, Shin MH, Lee YJ, Ahn YO. *Methodologic considerations on the cohort study of risk factors of stomach cancer: on the incompleteness of case ascertainment. Korean J Epidemiol 1997; 19: 152-60.*
13. Im JS, Kweon SS, Park KS, Sohn SJ, Choi JS. *Completeness estimation of the population-based cancer registration with capture-recapture methods. Korean J Prev Med 2000; 33: 31-5.*
14. Kim MH, Park JK, Ki MR, Hur YJ, Choi BY, Kim JS. *Evaluation of the completeness of case reporting during the 1998 Cheju-do mumps epidemic, using capture-recapture methods. Korean J Prev Med 2000; 33: 313-22.*
15. Ha M, Kwon HJ, Kang DH, Cho SH, Yoo KY, Joo YS, Sung JH, Kang JW,

Kim DS, Lee SI. *Completeness estimation of the Korean medical insurance data in childhood asthma. Korean J Prev Med 1997; 30: 428-36.*

16. Bae JM, Ahn YO. *A nested case-control study on the high-normal blood pressure as a risk factor of hypertension in Korean middle-aged men. J Korean Med Sci 2002; 17: 328-36.*
17. Bae J, Gwack J, Park SK, Shin HR, Chang SH, Yoo KY. *Cigarette smoking, alcohol consumption, tuberculosis and risk of lung cancer: the Korean multi-center cancer cohort study. J Prev Med Public Health 2007; 40: 321-8.*
18. Won YJ, Sung J, Jung KW, Kong HJ, Park S, Shin HR, Park EC, Ahn YO, Hwang IK, Lee DH, et al. *Nationwide cancer incidence in Korea, 2003-2005. Cancer Res Treat 2009; 41: 122-31.*

 http://dx.doi.org/10.3346/jkms.2012.27.5.489