RESEARCH ARTICLE

Estimating the Completeness of Lung Cancer Registry in Ardabil, Iran with a Three-Source Capture-Recapture Method

Mahmoud Khodadost[1,2]*, Alireza Mosavi-Jarrahi[3], Seyed Sepehr Hashemian[4], Fatemeh Sarvi[5], Khadije Maajani[6], Farhad Moradpour[7], Seyed Reza Khatibi[8], Hossein Amini[9]

Abstract

Cancer registration is an important component of a comprehensive cancer control program, providing timely data and information for research and administrative use. Capture-recapture methods have been used as tools to investigate completeness of cancer registry data. This study aimed to estimate the completeness of lung cancer cases registered in Ardabil Population Based Cancer Registry (APBCR) with a three-source capture-recapture method. Data for all new cases of lung cancer reported by three sources (pathology reports, death certificates, and medical records) to APBCR for 2006 and 2008 were obtained. Duplicate cases shared among the three sources were identified based on similarity of first name, last name and father's names. A log-linear model was used to estimate number of missed cases and to control for dependency among sources. A total of 218 new cases of lung cancer was reported by three sources after removing duplicates. The estimated completeness calculated by log-linear method was 26.4 for 2006 and 27.1 for 2008. The completeness differed according to gender. In men, the completeness was 26.0% for 2006 and 28.1 for 2008. In women, the completeness was 36.5% for 2006 and 46.9 for 2008. In conclusion, none of the three sources can be considered as a reliable source for accurate cancer incidence estimation.

Keywords: Capture-recapture - incidence estimation - log-linear - lung cancer - cancer registration - completeness

Asian Pac J Cancer Prev, **17**, Cancer Control in Western Asia Special Issue, 225-229

Introduction

Lung cancer is known as one of the most important public health problems because of its high incidence rate, rapid progression, and poor prognosis (Montazeri et al., 2001). Also, it is the leading cause of death due to cancer in 87 countries in men and 26 countries in women, with the latter largely restricted to high income countries(Torre et al., 2015). Lung cancer is one of the five leading cancers in Iran, and the incidence trend was increasing steadily in both men and women (Hosseini et al., 2009). Accurate cancer incidence data are essential for planning, monitoring and evaluating national and regional cancer control programs (Kamo et al., 2007). The purpose of population based cancer registries is to estimate the cancer burden in the area covered, to observe trends and regional differences and to provide a data base for epidemiological research (Schmidtmann, 2008). Decision makers in health authorities need to know how reliable the data is on which they base their policies. Therefore, completeness of registration is used as one of the measures of quality of a cancer registry (Schmidtmann, 2008). Completeness is defined as the proportion of incident cancer cases that is registered (Schmidtmann, 2008). Completeness level of cancer registration is one of the main parts of quality control in such registration (Mosavi-Jarrahi et al., 2013). In the literature, several methods are described to evaluate completeness, which are divided in two categories: qualitative methods and quantitative methods. The qualitative methods used were mortality/incidence (M:I) ratios and the proportion of microscopically verified cases and, among quantitative methods, the ones applied were the capture recapture, the death certificates and M:I ratios method, and the flow method (Castro, 2011). Since most cancer registries employ more than one data source for case finding,

*[1]Department of Epidemiology, Faculty of Health, [3]Dept. of Health & Community Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences, [2]Department of Epidemiology, Faculty of Health, Iran University of Medical Sciences, [4]Department of Psychology, Roozbeh Psychiatric Hospital, [6]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, [9]Research center for Prevention of Oral and Dental diseases, Baqiyatallah University of Medical Science, Tehran, Iran, [5]Department of Epidemiology and Biostatistics, School of Public Health, Hamadan University of Medical Science, Hamadan, [7]Social Determinants of Health Research Center, Kurdistan University of Medical Sciences, Sanandaj, [8]Torbat Heydariyeh University of Medical Sciences, Torbat Heydariyeh, Iran. *For correspondence: mahmodkhodadost@yahoo.com*

capture-recapture methods may be used to estimate the number of incident cases in the population, and hence to assess completeness of case ascertainment (Robles et al., 1988). Capture-recapture is the method widely used in wildlife population censuses (Suwanrungruang et al., 2011). Another important application for this method is in epidemiology for estimating prevalence of a particular disease and estimating the completeness of ascertainment of disease registers. However, capture-recapture method can principally be applied to any situation where there are two or even more incomplete lists (Poorolajal et al., 2010). Two assumptions have to be made when using the simple capture-recapture method. Firstly, the sources are independent, and secondly, all individuals within the same source have an equal chance of being included (Parkin and Bray, 2009; Suwanrungruang et al., 2011). The use of capture-recapture methods is very efficient for reducing the costs of disease registration as well as reducing bias in incidence estimations and in the case of comparing population subgroups (Mosavi-Jarrahi et al., 2013).

The pathology-based cancer registry has been established in Ardabil province since 1999. For the first time in Iran, the population based cancer registry established in Ardabil in 2003. The Ardabil cancer registry (ACR) actively collects information of cancer incidence from Pathology-based, hospital-based and death certificates. The main goal of the ACR is to measure cancer incidence and mortality in residents of Ardabil province (Babaei et al., 2010). As there is no screening test for lung cancer and the patients typically were detected in end stage of disease, and also because of its poor prognosis and rapid progression, the incidence cases may be not completely registered in pathology source and may registered only in death certificates or hospital records. So, this study aims to estimate the completeness of lung cancer in every source of registry (including pathology, hospital records and death certificates) and estimate the lung cancer incidence in Ardabil province by three source capture-recapture method.

Materials and Methods

This study was conducted in Ardabil Province which is located in north-west of Iran. All new cases of lung cancer reported by three sources; pathology reports, death certificates and medical records that reported to Ardabil population-based cancer registry in 2006 and 2008 were enrolled in this study. The duplicate cases in every 3 sources of registry were identified and removed using EXCEL software. Some characteristics such as name, surname, father's name, date of birth and ICD codes related to their cancer type were used to identify the common cases among three sources. After identifying the common cases using data linkage, the incidence rate of lung cancer was estimated by the capture-recapture method and log-linear models. To use capture-recapture method, two main assumptions should be considered, sources of information should be independently and all people who are in every data source should have an equal chance to presence in the study (Parkin and Bray, 2009; Suwanrungruang et al., 2011). However, in most human populations and medical Science studies, usually these assumptions are not established and different sources are not independent. So, three source capture-recapture and log-linear model was used to consider the interactions between three sources and estimate the completeness and more accurate incidence rate of lung cancer in Ardabil province. With three registers, there are eight possible combinations of these registers in which cases do or do not appear. The general model uses eight parameters, the common parameter (the logarithm of the number expected to be in all lists), three 'main effects' parameters (the log odds ratios against appearing in each list for cases who appear in the others), three 'two-way interactions' or second order effect parameters (the log odds ratios between pairs of lists for cases who appear in the other), and a 'three-way' interaction parameter. For three registers, A with i levels, B with j levels, C with k levels, the natural logarithm (ln or loge) of expected frequency

Table1. Model Selection in Log-Linear Analysis by AIC, BIC and G2 Statistics

Model	X***	N****	95% CI for N	DF**	G2*	BIC*	AIC*
P/C/D	202.4	420.4	(357.1-512.7)	4.0	70.9	104.3	104.5
PC/D	682.0	900	(601.2 -1431.8)	5.0	0.2	43.1	43.3
PD/C	116.4	334.4	(290.5 -404.8)	5.0	43.8	91.3	91.5
CD/P	118.2	336.2	(292.7 -405.0)	5.0	43.8	89.5	89.8
PC/PD	837.0	1055	(510.1 -2616.6)	6.0	0.0	44.8	45.1
PC/CD	620.0	838.0	(474.7 -1715.5)	6.0	0.2	44.9	45.3
PD/CD	42.4	260.4	(242.1 -292.3)	6.0	22.1	58.2	58.6
PC/PD/CD	820.6	1038.6	(375.4 -4495.4)	7.0	0.0	46.7	47.1

* Akaike's Information Criterion/ Bayesian Information Criterion/ Goodness of fit; ** Degree of freedom; *** The estimated number of lung cancer that were not recorded in any of three sources; **** The estimated total number of lung cancer in Ardabil province in 2006 and 2008; [a]D, death certificates source; P, pathology reports Source; C, hospital records; Model P/C/D, A model where all available resources are independent; Model PC/D, A model where sources P and C are dependent and independent of the source D; Model PD/C, A model where sources P and D are dependent and independent of the source C; Model CD/P, A model where sources C and D are dependent and independent of the source P; Model PC/PD, A model where two sources P and C and also two sources P and D are mutually interdependent and two sources C and D are independent; Model PC/CD, A model where two sources P and C and also two sources C and D are mutually interdependent and two sources P and D are independent; Model PD/CD, A model where two sources P and D and also two sources C and D are mutually interdependent and two sources P and C are independent; Model PC/PD/CD, A model where all two-way interaction between resources are exist

Fijk for cell ijk, ln Fijk, can be denoted as:

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_K^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

where θ is the common parameter , λA, λB, and λC are the main effect parameters, λAB, λAC and λBC are the second order effect (two-way interaction) parameters and λABC is the highest order effect (three-way interaction) parameter. The value of this last three-way interaction parameter cannot be tested from the study data and is assumed to be zero. To assess how the various log-linear models fit the data (model fitting) and select the best model we used the log likelihood-ratio test, also known as G2 or deviance, Akaike's Information Criterion (AIC) and Bayesian Information Criteria (BIC) which they can be expressed as $G^2 = -2\sum Obs_j \ln[Obs_j/Exp_{ji}]$

where Obsj is the observed number of individuals in each cell j, and Expji is the expected number of individuals in each cell j under model i. $AIC = G^2 - 2[df]$

where the first term, G2, is a measure of how well the model fits the data and the second term,2 [df], is a penalty for the addition of parameters (and hence model complexity). $BIC = G^2 - [\ln N_{obs}][df]$

Where Nobs is the total number of observed individuals. The lower the value of G2, AIC and BIC the better is the fit of the model. AIC is the more appropriate criteria which is used by researchers for model selection (Ho-ok and Regal, 1997; Hook and Regal, 2000; Motevalian et al., 2007). Therefore, we used these criteria for evaluating the goodness of fit. Finally the model with lower amount of the AIC was chosen as the best model.

Estimated incidence rate was calculated based on the estimated new cases of lung cancer (by use of selected model in log-linear analysis) at a certain time over the number of the population at risk in Ardabil province at that time. All the incidences are reported based on the incidence per one hundred thousand populations. Also the completeness was calculated by age groups and calendar time, respectively. In all stages of this study the individual's information such as name, surname and other characteristics were kept confidential. We used STATA software, version 12 (StataCorp, Texas, USA) for all computations.

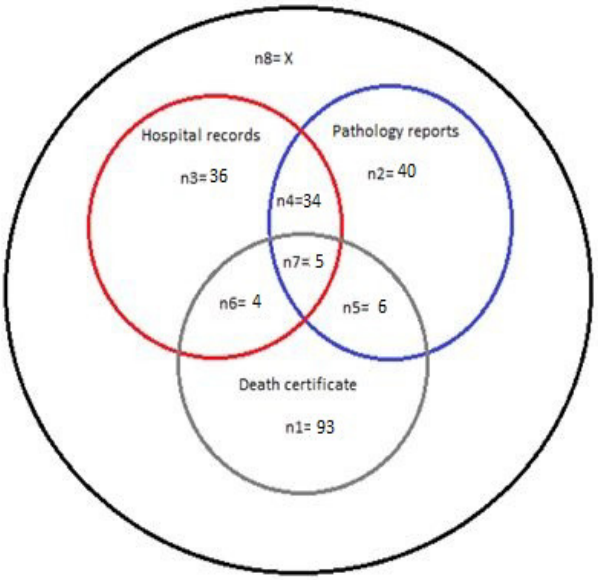

Figure 1. Venn Diagram of the Common Cases of Lung Cancer Between Pathology Reports, Hospital Records and Death Certificates

Results

After investigating and remove duplicate cases between three sources, total 218 new cases of lung cancer were reported to Ardabil population-based cancer registry in 2006 and 2008. The pathology source, hospital records and death certificates were reported 71, 81 and 113 new cases of lung cancer, respectively. Of 218 subjects 163 (74.8%) were male. The mean age of participants was 66.1 (±13.4), 66.3 (±13.3) years for men and 65.5 (±13.9) years for women. Venn diagram shows the common cases between pathology reports, hospital

Table 2. Estimated Incidence of Lung Cancer by Log-Linear Model Based on Ardabil Population in 2006 and 2008

Calendar year	Subgroups		Reported new cases*	Estimated new cases	95% CI for Estimated cases	"Completeness of registration** (%)"
	Gender	Male	64.0	264.0	(139.7 - 517.7)	26.1
2006		female	25.0	68.4	(43.5 - 147.8)	36.6
	Age groups	50.0≤	16.0	32.4	(24.5 - 65.1)	49.4
		51.0-64.0	18.0	64.0	(34.5 - 185.6)	28.2
		65.0≥	55.0	191.2	(111.6 - 398.7)	28.8
	Total		89.0	337.5	(203.5 - 638.6)	26.4
2008	Gender	Male	99.0	352.5	(212.1 - 667.6)	28.1
		female	30.0	64.3	(53.5 - 89.1)	46.9
	Age groups	50.0≥	20.0	52.5	(25.8 - 202.9)	38.1
		51.0-64.0	35.0	107.3	(61.1 - 265.9)	32.6
		65.0≤	74.0	230	(138.4 - 452.3)	32.2
	Total		129.0	477.4	(298.9 - 843.3)	27.1

*Number of new cases reported by pathology reports; Hospital reports and death certificates after removing duplicates; ** Number of registered cases divided by the number of estimated cases

records and death certificate (Figure 1). In three source capture-recapture analysis with log-linear model, a model in which two sources of pathology and medical records were mutually interdependent and death certificates source were independent was chosen as the best model with the lowest value of Akaike's Information Criterion and Bayesian Information Criterion that were 43.3 and 43.1, respectively (Table 1). The estimated total number of lung cancer in 2006 and 2008 was 900 (95%CI: 601.2-1431.8). The completeness of registration for all three sources after removing duplicates was 24.2% (218 cases) and also for pathology reports, hospital records and death certificates were 7.9% (71 cases), 9% (81 cases) and 12.5% (113 cases), respectively. The completeness of lung cancer was estimated generally and to the age and sex subgroups based on the estimated new cases of lung cancer (by use of selected model in log-linear analysis). The estimated completeness for 2006 and 2008 was 26.4% and 27%, respectively (Table 2).

Discussion

The completeness of lung cancer in this study was estimated by the capture-recapture method and log-linear models. The mean age of all subjects was 66.1±13.4 years (66.3 ± 13.3 for men and 65.5± 13.9 for women). The age distribution does not show the difference between men and women and is consistent with the average age reported by studies conducted in other parts of Iran that reported average age was 59-67 years (Montazeri et al., 2001; Hosseini et al., 2009). In log-linear analysis, we select the best model using AIC, BIC and G2 statistics. The model where sources pathology and hospital records are dependent and independent of the source death certificates was selected in this study. Also, the dependency between pathology reports and hospital records in selected model is seems logical with the description of this relationship in situations that what happens really in society. The completeness of registration based on population based cancer registry in Ardabil province, after removing duplicate cases between pathology reports, hospital records and death certificates, for 2006 and 2008 years was 26.4% and 27%, respectively. Also the completeness of registration in men and women, after removing the duplicates between three sources, in 2006 was 26.01% and 36.54%, and in 2008 was 28.08% and 46.87%, respectively. The estimated completeness in log-linear analysis for the completeness of cancer registries in our study is much lower than other countries that reported 96 % to 99.6% for all types of cancers overally (Robles et al., 1988; Crocetti et al., 2001; Gajalakshmi et al., 2001) and also for lung cancer, (Inger Kristin Larsen et al., 2009) reported the completeness of lung cancer registration in Norway as 96.91% during 2001 to 2005 years (Larsen et al., 2009) and Donna McClish et al reported the 74% completeness for lung cancer in Virginia Cancer Registry (McClish and Penberthy, 2004). There is no evidence for completeness of lung cancer registry in other parts of iran, but the similar study that conducted in Tehran with capture recapture method for evaluations of cancer registry system indicated that the completeness of gastric and esophagus cancer registry was 29.9% and 30.8%,respectively (Aghaei et al., 2013) and also in Ardabil province of Iran, the completeness of registration for gastric and esophagus cancer was 36.35% and 37.76% (Khodadost et al., 2014). This result is consistent with our study and confirmed that the quality of cancer registry in Iran is highly inappropriate and need to more attention to improve it.

Acknowledgements

The authors wish to express their sincere thanks to all staffs of the Ardabil cancer registry center, especially Dr. Babaei who sincerely helped the data collection.

References

Aghaei A, Ahmadi-Jouibari T, Baiki O, et al (2013). Estimation of the gastric cancer incidence in Tehran by two-source capture-recapture. *Asian Pac J Cancer Prev*, **14**, 673-7.

Babaei M, Pourfarzi F, Yazdanbod A, et al (2010). Gastric cancer in Ardabil, Iran--a review and update on cancer registry data. *Asian Pac J Cancer Prev*, **11**, 595-9.

Crocetti E, Miccinesi G, Paci E, et al (2001). An application of the two-source capture–recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy. *Eur J Cancer Prev,* **10,** 417-23.

Gajalakshmi V, Swaminathan R, Shanta V (2001). An independent survey to assess completeness of registration: population based cancer registry, Chennai, India. *Asian Pac J Cancer Prev*, **2,** 179-83.

Hook EB, Regal RR (1997). Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *Am J Epidemiol,***145**, 1138-44.

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,* **152,** 771-9.

Hosseini M, Naghan PA, Karimi S, et al (2009). Environmental risk factors for lung cancer in Iran: a case–control study. *Int J Epidemiol,* **38**, 989-96.

Kamo K-i, Kaneko S, Satoh K, et al (2007). A mathematical estimation of true cancer incidence using data from population-based cancer registries. *Jpn J Clin Oncol*, **37**, 150-5.

Khodadost M, Yavari P, Babaei M, et al (2014). Estimating the completeness of gastric cancer registration in Ardabil/ Iran by a capture-recapture method using population-based cancer registry data. *Asian Pac J Cancer Prev*, **16,** 1981-6.

Larsen IK, Småstuen M, Johannesen TB, et al (2009). Data quality at the cancer registry of norway: An overview of comparability, completeness, validity and timeliness. *Eur J Cancer,* **45**, 1218-31.

McClish D, Penberthy L (2004). Using medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. *Med Care,* **42**, 1111-6.

Montazeri A, Milroy R, Hole D, et al (2001). Quality of life in lung cancer patients: as an important prognostic factor. *Lung Cancer*, **31**, 233-40.

Mosavi-Jarrahi A, Ahmadi-Jouibari T, Najafi F, et al (2013). Estimation of esophageal cancer incidence in Tehran by log-linear method using population-based cancer registry data. *Asian Pac J Cancer Prev,* **14**, 5367-70.

Motevalian A, Holakoei naeini K, Mahmoodi M, et al (2007). Estimating deaths due to traffic accidents in Kerman using

capture-recapture method. *J Sch Public Health Inst Public Health Res,* **5**, 61-72.

Parkin DM, Bray F (2009). Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. *Eur J Cancer*, **45,** 756-64.

Poorolajal J, Haghdoost AA, Mahmoodi M, et al (2010). Capture-recapture method for assessing publication bias. *J Res Med Sci,* **15**, 107.

Robles SC, Marrett LD, Aileen Clarke E, et al (1988). An application of capture-recapture methods to the estimation of completeness of cancer registration. *J Clin Epidemiol,* **41,** 495-501.

Schmidtmann I (2008). Estimating Completeness in cancer registries–comparing capture-recapture methods in a simulation study. *Biom J,* **50**, 1077-92.

Suwanrungruang K, Sriplung H, Attasara P, et al (2011). quality of case ascertainment in cancer registries: a proposal for a virtual three-source capture-recapture technique. *sian Pac J Cancer Prev*, **12**, 173-8.

Torre LA, Bray F, Siegel RL, et al (2015). Global cancer statistics, 2012. *CA Cancer J Clin,* **65**, 87-108.