RESEARCH ARTICLE

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

Mahmoud Khodadost1,2,*, Alireza Mosavi-Jarrahí3, Seyed Sepehr Hashemian4, Fatemeh Sarvi5, Khadije Maajani5, Farhad Moradpour7, Seyed Reza Khatibi8, Hossein Amini9

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

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., 2001). 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-Jarrahí 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,
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

### Table 1. Model Selection in Log-Linear Analysis by AIC, BIC and G2 Statistics

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

* 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; ’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 PC/D, A model where sources P and D are dependent and independent of the source C; Model PC/D, 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/PD/CD, A model where two sources P and C 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
Estimating the Completeness of Lung Cancer Registration in Ardabil, Iran with a Three-Source Capture-Recapture Method


de Fin cell ijk, ln Fijk, can be denoted as:

\[ \ln F_{ijk} = \theta + \lambda_A i + \lambda_B j + \lambda_C k + \lambda_{AB} ij + \lambda_{AC} ik + \lambda_{BC} jk + \lambda_{ABC} ijk \]

where \( \theta \) is the common parameter, \( \lambda_A, \lambda_B, \) and \( \lambda_C \) are the main effect parameters, \( \lambda_{AB}, \lambda_{AC} \) and \( \lambda_{BC} \) are the second order effect (two-way interaction) parameters and \( \lambda_{ABC} \) 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:

\[ G^2 = -2 \sum \text{Obs}_{ij} \ln \left( \frac{\text{Obs}_{ij}}{\text{Exp}_{ij}} \right) \]

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). \[ \text{BIC} = G^2 - \ln N_{obs} [df] \]

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.

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

*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**


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