A Mathematical Estimation of True Cancer Incidence Using Data from Population-based Cancer Registries

Ken-ichi Kamo1,2, Satoshi Kaneko1,3, Kenichi Satoh4, Hirokazu Yanagihara5, Shoichi Mizuno6 and Tomotaka Sobue1

1Statistics and Cancer Control Division, Research Center for Cancer Prevention and Screening, National Cancer Center, Tokyo, 2Division of Mathematics, School of Medicine, Liberal Arts and Sciences, Sapporo Medical University, Sapporo, 3Nairobi Station for Research on Tropical Medicine, Institute of Tropical Medicine, Nagasaki University, Nagasaki, 4Department of Environmentmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima, 5Department of Mathematics, Graduate School of Science, Hiroshima University, Hiroshima and 6Epidemiology and Health Promotion, Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan

Received May 22, 2006; accepted September 24, 2006; published online February 1, 2007

Background: Accurate cancer incidence data are needed to plan, monitor and evaluate national cancer control programs. In Japan, however, such information is not available owing to incomplete cancer registries. In order to attain incidence estimation adjusted to account for this incomplete information, we have developed a new method.

Methods: We developed a nonlinear regression model between observed incidence/mortality ratios and proportions of death certificate notification to observed incidence in various cancer registries. This model enables us to obtain the ‘true incidence/mortality ratio’, which, in the regression curve, is at zero point for the proportion of death certificate notifications. This is an ideal registration state without any missing cases. By multiplying it by the number of cancer mortalities from the National Vital Statistics, corrected cancer incidence can be estimated.

Results: Applying this method for the estimation of the Japanese cancer incidence in 1997, we obtained the ‘true incidence/mortality ratios’ of 2.074 for men and 2.587 for women. Cancer incidences in Japan for 1997 were thus estimated to be 346,000 for men and 280,000 for women.

Conclusions: A new method is proposed to estimate the national cancer incidence after adjusting for completeness of cancer registries. This method enables us to more accurately estimate the cancer incidence in a country where several cancer registries exist with various degrees of completeness of registration.

Key words: neoplasms – cancer – incidence – registries – models (theoretical)

INTRODUCTION

Information about cancer incidence is fundamental for planning, monitoring and evaluating national and regional cancer control programs, as well as having accurate data on the number of cancer mortalities. In Japan, the mortality data can be obtained from the National Vital Statistics, which is essentially a summary of the death certificates issued. However, cancer incidence data are based on several prefecture-wide, voluntary-based cancer registries in Japan. The Research Group for Population-Based Cancer Registration has reported national estimates of cancer incidence since 1975 using selected population-based cancer registries (1,2). However, the estimation could be substantially subject to underestimation, because it is based on the incidence data from registries in which registration
completeness is not adequate compared with those registries of the USA or European countries (3), although the group selects only those data which fulfill their criteria of completeness. From the public health and policymaking points of view, accurate estimation of the cancer incidence needs to be presented, which is adjusted for cases currently unregistered.

In this paper, we introduce a new method to estimate the cancer incidence using data from several population-based cancer registries with various levels of completeness, and present as an example the estimation for the cancer incidence of Japan in 1997. The number obtained by the new method proposed in this paper is regarded as one adjusted only for quantitative completeness of reporting but not for qualitative aspects. We note that to ultimately estimate the accurate number of incidence we need information not only with quantitative but also qualitative sufficiency.

METHODS

Cancer incidence in population-based cancer registries is mainly determined by the cancer cases reported from hospitals. However, some newly diagnosed cancer cases are naturally not reported from hospitals. Such cases can be detected through their death certificates if the cancer was fatal. However, if the cancer was not fatal, they can not be detected because of the lack of reports from hospitals and also because those patients may die of other causes. Therefore, correction for such undetectable cases is critical for estimating the real incidence of cancer in a country where several population-based cancer registries co-exist with various degrees of completeness of registration.

In order to simplify the methodology, we divided newly diagnosed cancer cases into four groups according to registration and vital status:

(i) those who are already registered and who have died of cancer (a1);
(ii) those who are already registered and do not belong to a1, namely, individuals who survived or died of other causes (a2);
(iii) those who are not registered and died of cancer (a3);
(iv) those who are not registered and do not belong to a3 (a4).

Because the a4 cases are not detectable by the population-based cancer registries and are not included in the cancer incidence report, the reported number of cancer cases from these registries can be expressed as a1 + a2 + a3. The proportion of death certificate notification (DCN) cases to the observed incidence is calculated by adding a1 and the cases of cancer diagnoses and/or treatments in the death certificates as the numerator and adding a1 + a2 + a3 and the cases of cancer diagnoses and/or treatments in the death certificates as the denominator. Here we assume that the cases of cancer diagnoses and/or treatments in the death certificates are so small compared with a1 and a2 that we can approximate the proportion of DCN to \(a_1/(a_1 + a_2 + a_3)\) and incidence/mortality (IM) ratio to \((a_1 + a_2 + a_3)/(a_1 + a_3)\). Using a1, a2, a3 and a4, indicators needed for our estimation, i.e. degree of completeness of registration \(r\), proportion of DCN \(x\) and ‘true IM ratio’ \(K\), are expressed as below (Table 1).

The logic for our method of estimating the real figure of cancer incidence is as follows. We calculate an IM ratio \(y\) and a proportion of DCN \(x\) for every registry. If there is mathematical relation between the proportion of DCN \(x\) and the IM ratio \(y\), the ‘true IM ratio’ \(K\) is estimated as the value of the IM ratio at the zero point of the proportion of DCN. Using \(K\) and the number of mortalities, the real figure for mortality incidence can be estimated as \(K(a_1 + a_3)\). Thus, a critical step in our method is to obtain the mathematical relationship between the IM ratio \(y\) and the proportion of DCN \(x\).

Parkin et al. introduced an equation to estimate the degree of completeness of registration applied in a registry with low proportion of DCN as follows (4):

\[
r = \frac{1}{1 + (y - 1)x}
\]

(A)

Ajiki et al. modified this equation to apply to high proportion of DCN under the assumption that the ratio of a3 to a4 equals that of a1 to a2 (5):

\[
r = \frac{1 - xy}{1 - x}
\]

(ii)

The simple derivation of equations (I) and (ii) is described in Appendix 1. Using equation (ii) and the indicators in Table 1, the IM ratio \(y\) can be expressed as a dependent variable with only one independent variable, the proportion of DCN \(x\) as the following equation (III). Using the indicators and equations in Table 1, the unknown number a4 is represented as \(a_4 = (K - 1)(a_1 + a_3) - a_2\). Then the degree of completeness is estimated (1).

\[\text{Table 1. Indicators needed for estimation of the ‘true IM ratios’ from the various cancer registry data}\]

<table>
<thead>
<tr>
<th>Degree of completeness (r)</th>
<th>((a_1 + a_2 + a_3)/(a_1 + a_2 + a_3 + a_4))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of DCN (x)</td>
<td>(a_1/(a_1 + a_2 + a_3))</td>
</tr>
<tr>
<td>IM ratio (y)</td>
<td>((a_1 + a_2 + a_3)/(a_1 + a_3))</td>
</tr>
<tr>
<td>True IM ratio (K)</td>
<td>((a_1 + a_2 + a_3 + a_4)/(a_1 + a_3))</td>
</tr>
</tbody>
</table>

\(a_1\), already registered and died of cancer; \(a_2\), registered and not a1; \(a_3\), bit registered and died of cancer; \(a_4\), not registered and not a3 (unknown); degree of completeness, proportion of observed incidence to true incidence; proportion of DCN, proportion of DCN cases to observed incidence; IM ratio, a ratio of observed incidence to mortality; true IM ratio, IM ratio under complete registration.
completeness of registration \((r)\) can be expressed without \(a_4\) as:

\[
r = \frac{a_1 + a_2 + a_3}{a_1 + a_2 + a_3 + a_4} = \frac{a_1 + a_2 + a_3}{a_1 + a_2 + a_3 + (K - 1)(a_1 + a_3) - a_2} = \frac{v}{K}.
\]

Substituting this relation into equation (II), we obtain

\[
y = \frac{K}{1 + (K - 1)x}.
\]  

(III)

This equation means that the IM ratio \(y\) is expressed by the proportion of DCN \(x\) and a constant value, the ‘true IM ratio’ \(K\). In other words, the ‘true IM ratio’ \(K\) can be calculated from nonlinear regression model defined with the observed IM ratios \(y\) and proportion of DCN \(x\) from various population-based cancer registries. The relationships between DCN rates and IM ratios in equation (III) are shown in Fig. 1 with various ‘true IM ratios’ \(K\).

Under the assumption that the ‘true IM ratio’ \(K\) is uniform for every registry within a strata of sex and cancer sites, the observed IM ratio \(y_i\) and the proportion of DCN \(x_i\) at registry \(i\) have the probabilistic relationship from equation (III) as follows.

\[
E(y_i \mid x_i) = \frac{K}{1 + (K - 1)x_i},
\]

(IV)

where \(E(y_i \mid x_i)\) denotes the conditional expectation of \(y_i\) under given \(x_i\). We estimate the ‘true IM ratio’ \(K\) using the weighted maximum likelihood method. The validity of equation (III) and the detail to estimate the ‘true IM ratio’ are shown in Appendices 2 and 3, respectively.

RESULTS

As an example, we applied our method for estimating the real number of cancer incidence of all sites in Japan in 1997, using the numbers of mortality, observed incidence, DCN and population from eleven population-based cancer registries. These registries participate in the Research Group for Population-Based Cancer Registration in Japan. We plotted observed IM ratios and proportion of DCN for each cancer registry in Fig. 2 along with the most suitable regression curve for males and females using the maximum likelihood method, assuming that the random error term is independently distributed in an identical manner to a normal distribution with mean 0 (see Appendix 3).

Using these regression curves, we estimated the ‘true IM ratios’ \(K\) at the zero point of the proportion of DCN. The ‘true IM ratios’ were 2.074 for males and 2.587 for females (Table 2).

By multiplying the ‘true IM ratios’ by the number of cancer mortalities for the whole country, we obtained the real figure for cancer incidences in Japan in 1997 as about 346 000 for males and 280 000 for females, which are 26% and 37% larger than those reported by the Research group for population-based cancer registration. It is noted that the latter are widely used for cancer research and cancer-related policy making (1, 2, 6, 7).

DISCUSSION

For the purpose of estimating the real incidence of cancer using data from population-based registries with various
proportions of DCN (index of registration completeness), our method can provide a more accurate estimation compared with the current method (5–7) as it takes completeness of reporting into consideration. The current method calculates national cancer incidence simply using the arithmetic mean of observed incidence rate from selected cancer registries that fulfill the criteria for completeness. The selection criteria are: a proportion of death certificate only (DCO) to observed incidence of less than 0.25 or a proportion of DCN of less than 0.3, along with an observed IM ratio \( \geq 1.5 \) for all cancer sites of both sexes. Thus, the criteria are not so rigid and therefore the figure for cancer incidence must be an underestimation (8–10). An estimation using Poisson regression is employed in EU countries (11,12), but this method also does not correct for incomplete registrations.

In contrast, our method enables us to estimate cancer incidence compensating for incomplete registrations. The above examples of our method show results that are 1.26 and 1.37 times larger than the incidences currently reported for males and females, respectively, in 1997 for Japan. This means that there should have been newly diagnosed cases of about 71,000 males and 76,000 females in addition to the currently published cases of 275,276 males and 203,879 females (6).

We would like to emphasize that our method can include information from cancer registries with various levels of completeness of registration. However, this does not mean that a low degree of completeness is permissible. In principle, of course, it is ideal that a high degree of competence is maintained in all regions. As shown in Fig. 2, registry information is considered together as a whole for a regression model, along with regional population weight, which influences the estimate of national cancer incidence. This can reduce the influence of instability of incidence data from a region with a small population. The largest region had a population of about 15 times larger than the region with the smallest population in the 11 regions used for our estimate.

Our method has several limitations for cancer incidence estimation, which are related to the following assumptions. First, we assumed that the ‘true IM ratios’ are uniform in any regional registries. It would be natural, however, to expect that sex, age and site distribution are not uniform for every region. In addition, regional differences must exist for the IM ratio because of the disparity in cancer care quality and screening system. In order to simplify the model and calculations, we regarded these differences as random deviations from the National IM ratio in our model described in Appendix 3. Therefore our model may not work well, for example, if we want to investigate the regional differences.

In the future, it may be necessary to build the model taking regional differences into account, for example, as random effects.

Our second assumption is that cancer mortality rates among groups registered are equal to those not registered. In other words, the ratio \( a_3 \) to \( a_4 \) is equal to that of \( a_1 \) to \( a_2 \) in Table 1. This assumption is needed to estimate the unknown number \( a_4 \) from known numbers \( a_1 \), \( a_2 \) and \( a_3 \) in equation (II). This assumption is of course not valid if the cancer mortality rate differs between the registered and unregistered groups. If cancer mortality rates of a registered group were higher than those of an unregistered group, the unknown number \( a_4 \) or national number of incidence would be underestimated.

Third, we implicitly assume that the cancer mortality and incidence in the registries do not change throughout the periods for which estimation is made. This assumption is needed because our method is based on the IM ratios. Mortality cases are derived from the past incidences. Therefore, drastic change in the incidence and mortality rates will influence the IM ratios. Taking this assumption into account, the figures we have presented might still be underestimated, because cancer incidence might have been increasing and prognosis of cancer cases might have been improved.

Unfortunately, we can not verify these three assumptions at present. It may be natural to expect that these assumptions are not realistic. Hence we do not claim that we can ultimately know the true number of incidences. However it is very important to estimate the figure as close as possible to the true one in order to evaluate the trend of cancer incidence. For attaining a more accurate estimation to reduce the degree of underestimation in comparison to the figure currently reported, we temporally assume it until we can improve our method. The validation of these assumptions is our important next subject.

In conclusion, we presented a new method to compute the number of nationwide cancer incidences using the ‘true IM ratio.’ This method gives us a more accurate estimation regarding the national cancer incidence in a country where several cancer registries exist with various degrees of completeness. It should be noted that this method gives the figure adjusted only for the degree of completeness of reporting, i.e. the quantitative aspect, therefore qualitative aspects are not taken into account. Hence we can evaluate the quantitative aspect for cancer incidence, for example the extent of underestimation, using this method. From the viewpoint of

Table 2. Estimated ‘true IM ratio’ and the number of incidences for all cancer sites in Japan, 1997

<table>
<thead>
<tr>
<th></th>
<th>True IM ratio</th>
<th>Estimated incidence (a)</th>
<th>95% confidence interval</th>
<th>Currently reported incidence (b)</th>
<th>(a)/(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>2.074</td>
<td>346,425</td>
<td>(327,407, 367,061)</td>
<td>275,276</td>
<td>1.27</td>
</tr>
<tr>
<td>Female</td>
<td>2.587</td>
<td>280,241</td>
<td>(263,753, 297,989)</td>
<td>203,879</td>
<td>1.37</td>
</tr>
</tbody>
</table>

IM ratio, incidence/mortality ratio. Currently reported incidence is estimated by the Research Group for Population-based Cancer Registration.
Estimate the number of cancer incidence

cancer control, every registry should establish a system to collect complete and accurate cancer incidence data in its region, because both quantitative and qualitative aspects are essential for cancer registration.

Acknowledgements

The authors wish special thank to Professor Megu Ohtaki, Hiroshima University, Dr Wakiko Ajiki, National Cancer Center and Dr Hideaki Tsukuma, Osaka Medical Center for Cancer and Cardiovascular Diseases for their valuable advice and encouragement. We also thank the staff of all the registries participating in the Research Group for Population-Based Cancer Registration in Japan: Miyagi, Yamagata, Niigata, Fukui, Aichi, Shiga, Osaka, Tottori, Saga, Nagasaki and Okinawa prefectures, whose contribution to the collection and processing of data made this publication possible. This work was supported in part by a Grant-in-Aid for Cancer Research for the Ministry of Health, Labor and Welfare, Japan (8-2), and the Foundation for the Promotion of Cancer Research for the Third-Term Comprehensive 10-Year Strategy for Cancer Control. Ken-ichi Kamo’s research is supported by the Ministry of Education, Science, Sports and Culture Grant-in-Aid for Young Scientists (B), No. 18790398, 2006-2008.

Conflict of interest statement

None declared.

References


APPENDIX 1

THE DERIVATION OF PARKIN’S EQUATION (I) AND AJIKI’S EQUATION (II)

We assume that the ratio of $a_3$ to $a_4$ equals to that of $a_1$ to $a_2$, i.e. $a_4 = a_2 a_3 / a_1$.

Then the degree of completeness of registration $r$ can be denoted using the proportion of DCN $x$ and IM ratio $y$ by

$$ r = \frac{a_1 + a_2 + a_3}{a_1 + a_2 + a_3 + a_4} = \frac{a_1 + a_2 + a_3 + (a_2 a_3) / a_1}{1 + (a_2 / a_1) (a_3 / (a_1 + a_2 + a_3))} \quad (V) $$

Parkin considered the situation that the proportion of DCN $x$ is quite low. Then $(a_1 + a_2) / a_1$ can be approximated by IM ratio, $y = (a_1 + a_2 + a_3) / (a_1 + a_3)$. Substituting this equation into (I), we obtain Parkin’s equation (I).

Parkin’s equation (I) needs the assumption that the DCN rate is quite low. However, Ajiki et al. tried to expand Parkin’s equation (I) to the situation where a proportion of DCN is not so low. They noted that the IM ratio which does not include $a_3$, that is $(a_1 + a_2) / a_1$, can be written as

$$ \frac{a_1 + a_2}{a_1} = \frac{a_1 + a_2 + a_3 - a_3}{a_1 + a_3 - a_3} = \frac{1 - a_3 / (a_1 + a_2 + a_3)}{(a_1 + a_3) / (a_1 + a_2 + a_3) - a_3 / (a_1 + a_2 + a_3)} = \frac{1 - y}{1 - x} \frac{1}{(1/y) - x}. $$

Substituting this relationship into (V), we obtain Ajiki’s equation (II).

APPENDIX 2

THE VALIDITY OF EQUATION (III)

We validate equation (III) from two points of view. First, we note the fact that if all the incidence cases are detected as DCN cases, then the number of incidence is
equivalent to one of mortality, i.e. the IM ratio is equal to 1. Hence, the regression curve must pass through the point where both the proportion of DCN and IM ratio are equal to 1. We immediately note that for all $K > 1$ equation (III) satisfies this property (see Fig. 1).

Next, we consider the situation when the undetected cases occur one after another in a complete registry. Suppose the registration in a given region is complete, the relationship between the number of incidence $I$ and IM ratio $\kappa$, $a_1$, $a_2$, $a_3$ and $a_4$ can be expressed as $a_1 = I/K$, $a_2 = I(K - 1)/K$ and $a_3 = a_4 = 0$. If newly diagnosed cancer cases numbering $A$ are not reported, these relationships would become as follows: $a_1 = (I - A)/K$, $a_2 = (I - K)(K - 1)/K$, $a_3 = A/K$ and $a_4 = A(K - 1)/K$. Because the observed number of incidence is $a_1 + a_2 + a_3$ and the one of mortality is $a_1 + a_2$, the IM ratio $y$ and the proportion of DCN $x$ are expressed as

$$y = \frac{KI - A(K - 1)}{I} \text{ and } x = \frac{A}{KI - A(K - 1)},$$

respectively. After a simple calculation, we see that these satisfy equation (III). This implies that whether undetected cases increase or not, ‘the IM ratio – the proportion of DCN plots’ lie on the line of equation (III).

APPENDIX 3

THE STATISTICAL MODEL AND ESTIMATE OF THE UNKNOWN PARAMETERS

For area $i$, let $x_i$ and $y_i$ be the observed proportion of DCN and IM ratio, respectively. Because $y_i \in (1, \infty)$ we consider the next model that is equivalent to (IV):

$$\tilde{y}_i = \log \left( \frac{K}{1 + (K - 1)x_i} - 1 \right) + \epsilon_i,$$

where $\tilde{y}_i = \log(y_i - 1)$, and the random error term $\epsilon_i$ is assumed identically independently distributed according to a normal distribution with mean 0 and variance $\sigma^2$. The two unknown parameters in this model are $K$ and $\sigma^2$. These are estimated by maximizing the population weighted log-likelihood function. The details for estimation (especially for $\sigma^2$) are seen in (13).