Title: A mathematical estimation of cancer incidence using data from population-based cancer registries

Authors:

Ken-ichi Kamo[1,2],

Satoshi Kaneko[1],

Kenichi Satoh[3],

Hirokazu Yanagihara[4],

Shoichi Mizuno[5] and

Tomotaka Sobue[1]

Affiliations:

[1] Statistics and Cancer Control Division, Research Center for Cancer Prevention and Screening, National Cancer Center

[2] Division of Mathematics, School of Medicine Liberal Arts and Sciences, Sapporo Medical University

[3] Department of Environmentrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University

[4] Department of Social Systems and Management, Graduate School of Systems and Information Engineering, University of Tsukuba

[5] Epidemiology and Health Promotion, Tokyo Metropolitan Institute of Gerontology

Corresponding author:

Ken-ichi Kamo

Division of Mathematics, School of Medicine Liberal Arts and Sciences, Sapporo Medical University, S1W16, Chuoku, Sapporo 060-8543, Japan

Phone: +81-11-611-2111 (ext. 2594).

email: kamo@sapmed.ac.jp

Abstract

Accurate cancer incidence data is needed to plan, monitor and evaluate national cancer control programs. In Japan, however, such information is not available due to low reporting rates to cancer registries. In order to have more accurate incidence estimation, we developed a new method. We obtained a nonlinear regression model between observed incidence/mortality (IM) ratios and death certificate notification (DCN) rates in various cancer registries. This model gives us the ”true IM ratio”, which is at zero point of the DCN rate in the regression curve. This is an ideal registration state without any missing cases. By multiplying it to cancer mortality number from the National Vital Statistics, corrected cancer incidence can be estimated. The results of applying this method for Japanese cancer incidence estimation in 1997 were the true IM ratios 2.074 for men and 2.587 for women. Cancer incidences in Japan were estimated to be 346,000 for men and 280,000 for women. This method gives us the best guess regarding the National cancer incidence in a country having several cancer registries with various qualities. However, from the viewpoint of cancer control, every registry should establish a system to collect cancer incidence data completely in a region.

Key words

Neoplasms, Cancer, Incidence, Registries, Models(Theoretical)

Introduction

Knowing cancer incidence number is fundamental for planning, monitoring and evaluating national and regional cancer control programs, as well as having accurate data on cancer mortality number. In Japan, mortality data can be obtained from the National Vital Statistics, which are summarized from death certificates. On the other hand, cancer incidence data are derived only from 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 should be subject substantially to underestimation, because the estimation is based on incidence data from registries in which the registration completeness is not adequate level compared with that in cancer registries of U.S. or European countries [3], thought the Group selects that fulfill criteria for completeness.

From the public health and political point of view, accurate estimation on cancer incidence numbers, which include currently undetected cases, need to be presented.

In this paper, we introduced a new method to estimate the cancer incidence number using data from several population-based cancer registries with various completeness levels and illustrate an example of the estimate for cancer incidence number of Japan in 1997.

Methods

Cancer incidence in population-based cancer registries is mainly determined by cancer cases reported from hospitals. However, some newly diagnosed cancer cases are not reported from hospitals. Such cases can be detected by their death certificates if the cancer is fatal. On the other hand, if the cancer is not fatal, such cases can not be detected without reporting from hospitals. Therefore, correction of such undetectable cases is critical to estimate the national cancer incidence number in a country having several population-based cancer registries with various reporting rates from hospital.

In order to simply our methodology explanation, we devided newly diagnosed cancer cases into four groups according to registration and vital status;

1) those who are already registered and who have died of cancer (a_1);

2) those who are already registered and do not belong to a_1, namely, individuals who survived or died of other causes (a_2);

3) those who are not registered and died of cancer (a_3);

4) those who are not registered and do not belong to a_3 (a_4).

Herein, we defended the cases of a_3 as death certificate notification (DCN) cases including follow-back cases, which are triggered by the death certificate and registered cases by tracing back to the hospital regarding cancer diagnoses.

Since the a_4 cases are not detectable by the population-based cancer registries and they are not included in the cancer incidence report, the reported number of incidence cases by cancer registries can be expressed as $a_1+a_2+a_3$. The DCN rate is calculated with a_3 as numerator and $a_1+a_2+a_3$ as denominator, and cancer mortality cases is expressed as a_1+a_3. Using a_1, a_2, a_3 and a_4, indicators needed for our estimation, i.e., registration rate (r), DCN rate

(x), incidence/mortality (IM) ratio (y), and true IM ratio (K), are expressed (Table 1).

The logic for our method to estimate cancer incidence number is as follows. We calculate an IM ratio (y) and a DCN rate (x) for every registry. If there is methematiocal relation between the DCN rate (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 DCN rate. Using the K and mortality number, incidence number can be estimated as $K(a_1 + a_3)$. Thus, a critical step in our method is to obtain the mathematical relation between the IM ratio (y) and the DCN rate (x).

Parkin *et al* introduced an equation to estimate the registration rate applied in a registry with low DCN rates [4];

$$r = \frac{1}{1 + (y - 1)x}. \tag{1}$$

Ajiki *et al* modified this equation to apply to high DCN rates under the assumption that the ratio of a_3 to a_4 equals that of a_1 to a_2 [5]:

$$r = \frac{1 - xy}{1 - x}. \tag{2}$$

The derivation of equation (1) and (2) is described in Appendix 1 and 2, respectively. Using equation (2) and the indicators in Table 1, the IM ratio

(y) can be expressed as a dependent variable with only one independent variable, the DCN rate (x) as following equation (3). Using the indicators and equations in Table 1, the unknown number a_4 is represented as $a_4 = (K-1)(a_1+a_3) - a_2$. Then the registration rate (r) can be expressed without a_4 as

$$r = \frac{a_1+a_2+a_3}{a_1+a_2+a_3+a_4} = \frac{a_1+a_2+a_3}{a_1+a_2+a_3+(K-1)(a_1+a_3)-a_2} = \frac{y}{K}.$$

Substituting this relation into equiation (2), we obtain

$$y = \frac{K}{1+(K-1)x}. \tag{3}$$

This equation means that the IM ratio (y) is explained by the DCN rate (x) and a constant value, the true IM ratio (K). In other words, the true IM ratio (K) can be detected from nonlinear regression model defined with the observed IM ratios (y) and DCN rates (x) from various population-based cancer registries. The relationships between DCN rates and IM ratios in equation (3) are shown in Figure 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 and the DCN rate x at registry i have the probabilistic relationship from equation

(3) as follow.

$$E(y_i|x_i) = \frac{K}{1 + (K-1)x_i}, \tag{4}$$

where $E(y_i|x_i)$ 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 (3) and the detail to estimate the true IM ratio are shown in Appendix 3 and 4, respectively.

Results

As an example, we applied our method for estimating cancer incidence number of all sites in Japan in 1997, using mortality, observed incidence, DCN and population numbers from 11 population-based cancer registries in Japan. These registries participate in the Research Group for Population-Based Cancer Registration in Japan. We plotted observed IM ratios and DCN rates of each cancer registry in Figure 2 along with the most suitable regression curve for males and females using the maximum likelihood method, assuming that the random error term is identically independently distributed a normal distribution with mean 0 (see Appendix 4). Using these regression curves, we estimated the true IM ratios (K) at the zero point of the DCN rate. The true IM ratios were 2.074 for male and 2.587 for female (Table 2). By multiplying the true IM ratios by the cancer mortality numbers of the whole country, we obtained the cancer incidence numbers in Japan in 1997 as about 346,000 for males and 280,000 for females, which numbers were 26% and 37% larger than those of currently reported incidence cases by the Research group for population-based cancer registration, which are used broadly for cancer research and cancer-related policy making [1, 2, 6, 7].

Discussion

With regard to cancer incidence estimation using data from population-based cancer registries with various DCN rates (registration completeness), our method gives us more accurate estimation on cancer incidence in Japan compared with the current estimation method [5, 6, 7]. The current method calculates national cancer incidence using an arithmetic mean of observed incidence rate from selected cancer registries that fulfill the criteria for completeness. The selection criteria is; a death certificate only (DCO) rate of less than 0.25 or a DCN rate of less than 0.3, along with an observed IM ratio equal to or more than 1.5 for all cancer sites of both sexes. Thus, the selection criteria is not so rigid, that the estimated numbers of cancer incidence must be underestimated [8, 9, 10]. Estimation using Poisson regression is used in EU counties [11, 12], but this method can not correct incompletes of registration, either.

On the other hand, our method enables us to estimate cancer incidence compensating imcompleteness of registration. The example results of our method showed 1.26 and 1.37 times larger than currently reported incidences for males and females in 1997 in Japan. The additional new diagnosed cases were about 71,000 males and 76,000 females from the currently published cases of 275,276 males and 203,879 females [6].

Furthermore, characteristic of our method is usage of information from cancer registries with various completeness of registration. As shown Figure 1, registry information are considered together as a whole for a regression model along with regional population weight, which influenced the estimate of national cancer incidence. This can lessen influences of unstability on incidence data from a region with small population. The largest region had about 15 times more residents than a region with the smallest population size in 11 regions used for our estimate.

Although our method has several demerits for cancer incidence estimation, it depends on following assumptions. First assumption is that the true IM ratios be uniform in any regions where a registry exists. It is naturally expected that sex, age and site distribution are not similar for every region. Additionally there must exists a regional difference of the IM ratio caused by disparities in cancer care quality and early detection rates. Therefore, our model is not appropriate to investigate the regional differences.

Second assumption is that cancer mortality rates among already registered groups be equal to ones not registered; i.e., the ratio a_3 to a_4 be 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 (2). This assumption is not valid when 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 unregisterd group, an unknown number a_4, or national incidence number would be underestimated.

Thirdly, we implicitly assume that the cancer mortality, incidence and follow-back status in the registries have not changed throughout periods from past to present. This assumption is needed because our method is based on

IM ratios. Mortality cases are the past incident cases, therefore, changing incidence and mortality rates have influence on the IM ratios. Furthermore, since the currently reported incidence consists of regularly reported cases, follow-back cases, and DCO cases which are equally dealt as same incidence cases in a registry, changes in follow-back status has influense on the count of incidence number. As a result, the IM ratios are also affected. Taking this assumption into account, the figures we presented in this paper might still be underestimated because cancer incidence might have been increasing and prognosis of cancer cases might have been improved.

Furthermore, we do not stratify the estimation by age group. Since it is reasonable to expect that the DCN rates and IM ratios differ among age groups, we need to develop a method to estimate cancer incidence stratified by age group.

In conclusion, we presented a new method to compute nation-wide cancer incidence number using the true IM ratio. This method gives us more accurate estimation regarding the National cancer incidence in a country having several cancer registries with various qualities; however, from the viewpoint of cancer control, every registry should establish a system to collect cancer incidence data completely in a region.

Acknowledgements

The authors wish special thank to Prof. Megu Ohtaki, Hiroshima University, Dr. Wakiko Ajiki 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.

References

1. Research Group for Population-based Cancer Registration in Japan. Cancer incidence in Japan, 1975 - cancer registry statistics. GANN Monogr Cancer Res 1981; **26**: 92-116.

2. Research Group for Population-based Cancer Registration in Japan. Cancer incidence in Japan, 1985-89: re-estimation based on data from eight population-based cancer registries. Jpn J Clin Oncol 1998; **28**: 54-67.

3. IARC Press, Lyon. Cancer Incidence in Five Continents, Vol. VIII. IARC Scientific Publication 2002; **155**.

4. Parkin D, Chen V, Ferlay J, et al. Comparability and quality control in cancer registration. Lyon: IARC Press 1994.

5. Ajiki W, Tsukuma H, Oshima A. Index for evaluating completeness of registration in population-based cancer registries and estimation of registration rate at the Osaka Cancer Registry between 1966 and 1992 using this index. Nippon Koshu Eisei Zasshi 1998; **45**: 1011-7 (in Japanese).

6. Research Group for Population-based Cancer Registration in Japan. Cancer incidence and incidence rates in Japan in 1997: estimates based on data from 12 population-based cancer registries. Jpn J Clin Oncol 2002; **32**: 318-22.

7. Research Group for Population-based Cancer Registration in Japan. Cancer incidence and incidence rates in Japan in 1999: estimates based on data from 11 population-based cancer registries. Jpn J Clin Oncol 2004; **34**: 352-6.

8. Kato I, Tominaga S, Ikari A. Estimation of trends in cancer incidence in a population-based cancer registry. Nippon Koshu Eisei Zasshi 1990; **37**: 861-6 (in Japanese).

9. Inoue M, Tajima K, Inuzuka K, et al. The estimation of cancer incidence in Aichi Prefecture, Japan: use of degree of completeness of registration. J Epidemiol 1998; **8**: 60-4.

10. Brenner H. Limitations of the death certificate only index as a measure of incompleteness of cancer registration. Br J Cancer 1995; **72**: 506-10.

11. Black RJ, Bray F, Ferlay J, et al. Cancer incidence and mortality in the European Union: cancer registry data and estimates of national incidence for 1990. Eur J Cancer 1997; **33**: 1075-107.

12. Jensen OM, Esteve J, Moller H, et al. Cancer in the European Community and its member states. Eur J Cancer 1990; **26**: 1167-256.

13. Seber GAF, Wild CJ. Nonlinear regression. Wiley series in probability and statistics 1998.

Appendix 1: The derivation of Parkin's equation (1)

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

$$a_4 = \frac{a_2 a_3}{a_1}.$$

Then the registration rate r can be denoted using the DCN rate x and IM ratio y by

$$\begin{aligned} r &= \frac{a_1 + a_2 + a_3}{a_1 + a_2 + a_3 + a_4} = \frac{a_1 + a_2 + a_3}{a_1 + a_2 + a_3 + (a_2 a_3)/a_1} \\ &= \frac{1}{1 + (a_2/a_1)\{a_3/(a_1 + a_2 + a_3)\}} = \frac{1}{1 + (a_2/a_1)x} \\ &= \frac{1}{1 - x + \{(a_1 + a_2)/a_1\}x}. \end{aligned} \tag{5}$$

If the DCN rate x is quite low, then we can approximate $(a_1 + a_2)/a_1$ by IM ratio $y = (a_1 + a_2 + a_3)/(a_1 + a_3)$. Hence we obtain equation (1).

Appendix 2: The derivation of Ajiki's equation (2)

Parkin's equation (1) needs the assumption that the DCN rate is quite low. On the other hand, Ajiki *et al* tried to express the resistration rate r by DCN rate x and IM ratio y without this assumption. They noted that the IM ratio which does not include DCN case, that is $(a_1 + a_2)/a_1$, can be written as

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

Substituting this relationship into (5), we immediately obtain equation (2).

Appendix 3: The validity of equation (3)

First, we note the fact that if all the incidence cases are detected as DCN cases, then the incidence number is equivalent to the mortality number, i.e., the IM ratio is equal to 1. Hence, the regression curve must pass through the point where both the DCN rate and IM ratio are equal to 1. We immediately note that for all $K > 1$ equation (3) satisfies this (Figure 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; that is, the DCN (or DCO) rate (x) is equal to zero, the relationship between incidence number (I) and IM ratio (K), 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 to the registry, these relationships would become as follows: $a_1 = (I - A)/K$, $a_2 = (I - K)(K - 1)/K$, $a_3 = A/K$ and $a_4 = A(K-1)/K$. Then the observed incidence number is $a_1 + a_2 + a_3 = [(I - A)(K - 1) + I]/K$. Since the observed mortality, which is I/K, is invariable, the expected values of the IM ratio (y) and DCN rate (x) are

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

respectively. After a simple calculation, we see that these satisfy equation (3). This implies that whether undetected cases increase or not, the IM ratio - DCN rate plots lie on the line of equation (3).

Appendix 4: The estimate of the true IM ratio and its confidence

interval

For area i, let x_i and y_i be the observed DCN rate and IM ratio, respectively. Since $y_i \in (1, \infty)$, we consider the next model that is equivalent to (4):

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

where $\tilde{y}_i = \log(y_i - 1)$, and the random error term ε_i is assumed identically independently distributed according to a normal distribution with mean 0 and variance σ^2. Noting that $K > 1$, we transform K to $1 + \exp(\theta)$, where θ is a real number. Substituting this into above equation yields

$$\tilde{y}_i = \log\left(\frac{1 - x_i}{e^\theta + x_i}\right) + \varepsilon_i.$$

Let the population weight be w_i (satisfying $\sum_{i=1}^N w_i = 1$, where N is the number of area), w_i proportional to the population size of area i, and the expected value of $\tilde{y}_i$ be μ_i. Then the logarithmic likelihood function ℓ with population weight is

$$\ell(K, \sigma^2) = -\frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^N w_i(\tilde{y}_i - \mu_i)^2\right].$$

We will estimate $\hat{K} = 1 + \exp(\hat{\theta})$ and $\hat{\sigma}^2$, the estimators of K and σ^2, a view point of attaining the maximum of ℓ.

The confidence interval of parameter K is obtained from the assumption that $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically distributed according to a normal distribution with mean 0 and variance $\tau\sigma^2$ [13], where

$$\tau = \tau(\theta) = \frac{N \sum_{i=1}^{N} \frac{w_i^2}{(1+x_i \exp(\theta))^2}}{\left[\sum_{i=1}^{N} \frac{w_i}{(1+x_i \exp(\theta))^2}\right]^2}.$$

Hence $K - 1$ is asymptotically distributed to a log-normal distribution. Let $\hat{\tau}$ be a plug-in estimator of τ which is $\hat{\tau} = \tau(\hat{\theta})$. We obtain the asymptotically $1 - \alpha$ confidence interval of K as

$$1 + \exp\left[\hat{\theta} - z_{\alpha/2}\hat{\sigma}\sqrt{\frac{\hat{\tau}}{N}} - \frac{\hat{\tau}\hat{\sigma}^2}{2N}\right] \leq K \leq 1 + \exp\left[\hat{\theta} + z_{\alpha/2}\hat{\sigma}\sqrt{\frac{\hat{\tau}}{N}} - \frac{\hat{\tau}\hat{\sigma}^2}{2N}\right],$$

where z_α is an upper $\alpha \times 100$ percentage point of the standard normal distribution.

Table 1.

Indicators needed for estimation of the true Incidence/Mortality (IM) ratios from the various cancer registry data

Registration rate (r)	$(a_1 + a_2 + a_3)/(a_1 + a_2 + a_3 + a_4)$
DCN rate (x)	$a_3/(a_1 + a_2 + a_3)$
IM ratio (y)	$(a_1 + a_2 + a_3)/(a_1 + a_3)$
True IM ratio (K)	$(a_1 + a_2 + a_3 + a_4)/(a_1 + a_3)$

a_1 : already registered and died of cancer,

a_2 : registered and not a_1,

a_3 : not registered and died of cancer,

a_4 : not registered and not a_3 (unknown),

Registration rate : proportion of observed incidence to true incidence,

DCN rate : proportion of DCN cases to observed incidence,

IM ratio : a ratio of observed incidence to mortality,

True IM ratio : IM ratio under complete registration.

Table 2.

Estimated true incidence/mortality (IM) ratio and incidence number for all cancer sites in Japan, 1997

	True IM ratio	95% confidence interval	Estimated incidence (a)	95% confidence interval	Currently reported incidence (b)	(a)/(b)
Male	2.074	(1.96, 2.20)	346516	(327469, 367568)	275276	1.26
Female	2.587	(2.28, 2.94)	280268	(247008, 318511)	203879	1.37

Currently reported incidence is estimated by Reserch Group for Population-Based Cancer Registration.

Figure legends

Figure 1. Examples of the relationship between DCN rates and incidence/mortality (IM) ratios with different true IM ratios (1, 1.2, 1.6, 2, 3, 4 and 6).

Figure 2. Regression curves for the estimate of the true incidence/mortality (IM) ratios for all cancer sites. The size of the plot is proportional to the population size covered by the registries. The line denotes the regression curve.

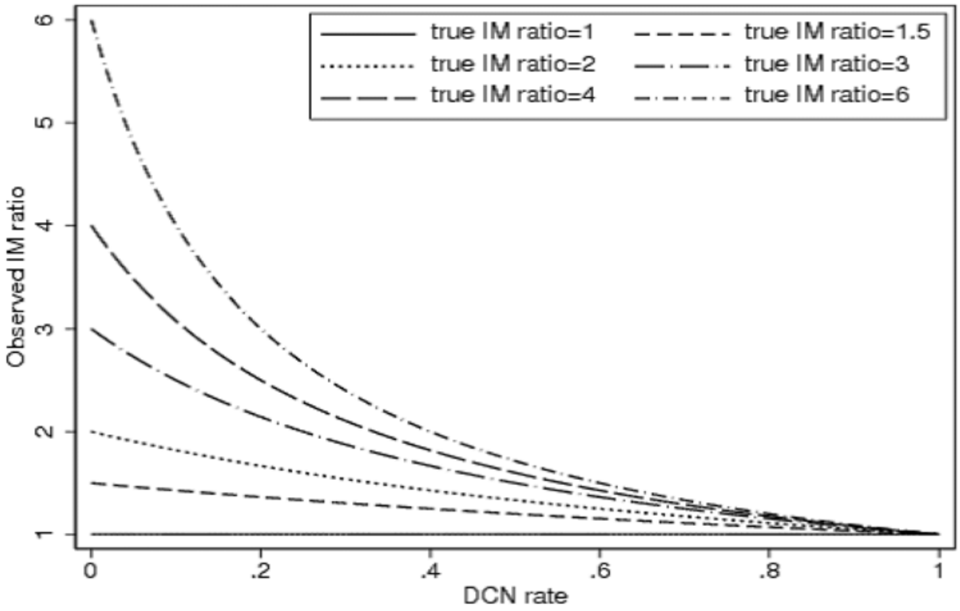

Figure 1

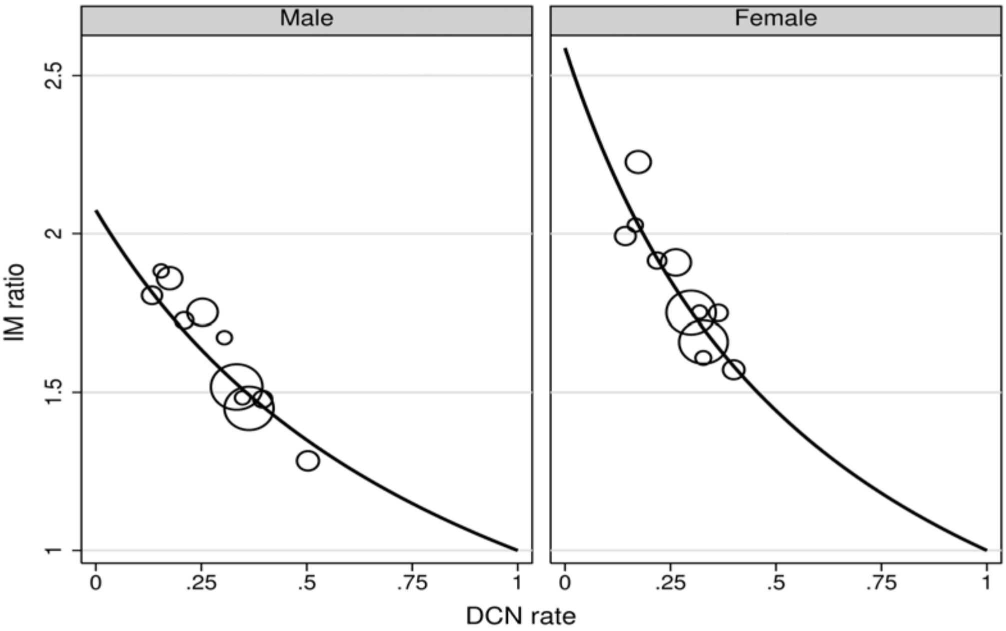

Figure 2