Cancer Epidemiology 39 (2015) 256–264

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Correction of misclassification bias induced by the residential mobility in studies examining the link between socioeconomic environment and cancer incidence

Josephine Bryere [a,*], Carole Pornet [a,b], Olivier Dejardin [a,b], Ludivine Launay [a], Lydia Guittet [a,b], Guy Launoy [a,b]

[a] U1086 INSERM, Cancers & Preventions, Avenue du General Harris, 14076 Caen, France
[b] CHU, Avenue de la Cote de Nacre, 14000 Caen, France

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 5 December 2014
Accepted 17 December 2014
Available online 9 January 2015

Keywords:
Cancer incidence
Bias (epidemiology)
Residential mobility
Misclassification bias
Poisson regression
Socioeconomic status

ABSTRACT

Background: Many international ecological studies that examine the link between social environment and cancer incidence use a deprivation index based on the subjects' address at the time of diagnosis to evaluate socioeconomic status. Thus, social past details are ignored, which leads to misclassification bias in the estimations. The objectives of this study were to include the latency delay in such estimations and to observe the effects.
Methods: We adapted a previous methodology to correct estimates of the influence of socioeconomic environment on cancer incidence considering the latency delay in measuring socioeconomic status. We implemented this method using French data. We evaluated the misclassification due to social mobility with census data and corrected the relative risks.
Results: Inclusion of misclassification affected the values of relative risks, and the corrected values showed a greater departure from the value 1 than the uncorrected ones. For cancer of lung, colon–rectum, lips–mouth–pharynx, kidney and esophagus in men, the over incidence in the deprived categories was augmented by the correction.
Conclusions: By not taking into account the latency period in measuring socioeconomic status, the burden of cancer associated with social inequality may be underestimated.

1. Introduction

The issue of social inequality in cancer incidence is of great importance for public health, and it has been well documented and debated in numerous countries [1]. Accurately assessing the influence of socioeconomic status on the risk of developing cancer largely depends on correctly evaluating an individual's social environment. Among recent studies, few have analyzed the link between cancer incidence and such individual indicators as financial resources, education, and profession; the great majority of investigations have used deprivation indices measured at an aggregate level which reflect the social environment in its entire individual and collective dimension [2,3]. In general, the place of residence or the exact address of patients at the time of diagnosis is used to assign the deprivation index.

The main limitation with such an approach is that in the vast majority of studies based on cancer registries, the address used to assess the patient's environment is that at the time of diagnosis; patients may have lived in another or several other places in the years before the diagnosis was made [4–8]. Indeed, the social environment considered in those studies is that at the time of cancer diagnosis. However, in the development of subclinical cancer, the time that should be evaluated is that preceding the onset of cancer, owing to the latency between exposure to risk factors and the cancer diagnosis [9,10]. Failure to make such evaluations may result in misclassifying the exposure of interest (socioeconomic environment) and thus affect the validity of the results.

The objective of this paper is to apply a method for correcting the above-described bias in estimating the relationship between social environment and cancer risk. We applied this method by

* Corresponding author. Tel.: +33 231 45 86 00; fax: +33 231458630.
E-mail address: josephine.bryere@inserm.fr (J. Bryere).

http://dx.doi.org/10.1016/j.canep.2014.12.008
1877-7821/

making use of French census-based data, which showed the changes in residence of patients diagnosed with cancer over a 10-year period. We estimated the proportion of individuals whose socioeconomic level altered as reflected in their changing place of residence over the 10-year period; we investigated the impact of misclassification on assessing the relationship between social environment and cancer risk.

2. Theory

The relationship between socioeconomic environment and cancer incidence can be modeled as follows. Let K be the number of areas: in studies addressing the issue of social inequality in cancer using an aggregated approach, the size of geographic areas has to be as small as possible. Let Y be the response variable. Y_k represents the observed number of cancer cases in area k. E_k is the expected number of cases in area k, taking into account the age structure of the population: $E_k = \sum_j t_j P_j$, where t_j is the global incidence rate for the age group j and P_j is the population size for the area k and the age group j. Let X be the exposure variable. In studies addressing the issue of social inequality in cancer using an aggregated approach, the socioeconomic environment of people living in a given area is usually assessed with an index based on available census data [11–15]. Such indices can be used with percentiles. In this paper, we used quintiles ranging from 1 to 5: fifth of the population whose deprivation index is less than the quintile 1 is the most affluent population, fifth of the population whose deprivation index is greater than the quintile 5 is the most deprived population. For the sake of clarity, the fifths are considered categories in the following. $\beta = (\beta_1, \ldots, \beta_5)$ is the regression coefficient that corresponds to X. The model, which is a Poisson regression, is written as follows:

$$\log(Y_k) = \log(E_k) + \alpha + \beta X \quad (1)$$

In this model, α is the intercept, β estimates the effect of socioeconomic environment on cancer incidence for four of the five categories (the lowest category is considered as the reference category). Our aim was to investigate the bias in calculating β when the socioeconomic environment was assessed by residence at the time of diagnosis-without taking into account possible changes of residence in the years preceding diagnosis.

To correct the regression coefficients, we implemented the methodology of Veierod and Laake [16]. Their method provides an expression that connects the naive coefficients (coefficients produced by simple analysis using only the address at the time of diagnosis) and the corrected coefficients (coefficients based on the fact that not all people at a given socioeconomic level occupied the same level several years earlier) in Poisson regressions.

Using such a method supposes that the bias corrected is non-differential. We may reasonably make this supposition: the probability of having moved does not differ between individuals diagnosed or not diagnosed with cancer; the period of interest concerns the years before the cancer is diagnosed, and cancer diagnosis does not affect the probability of having moved several years earlier.

Following the notations of Veierod and Laake [16], let $i = 1, 2, \ldots$, 5 and $j = 1, 2, \ldots, 5$ index the categories of the corrected and observed socioeconomic status, respectively. The lowest category (i, $j = 1$) was chosen as the reference category. Let X_i be a design variable indexing the exposure category several years before i for $i = 2, 3, \ldots, 5$ and W_j the design variable for the observed exposure category (socioeconomic level at the time of diagnosis).

Let $\pi_{i/j}$ denote the misclassification probability, i.e., the probability of having been at level i several years earlier but being observed at level j at the time of diagnosis.

Let $\beta_{1x}, \beta_{2x}, \ldots, \beta_{5x}$ be the corrected coefficients. The regression of Y on W is the observed regression model. The coefficients in this regression model, $\beta_{1w}, \beta_{2w}, \ldots, \beta_{5w}$ are the naive coefficients, and they are typically biased. The expressions of the naive coefficients explained by the corrected coefficients, given in the article [16], are as follows:

$$\beta_{jw} = \ln\left(\pi_{1/j} + \sum_{i=1}^{5} \pi_{i/j} \exp(\beta_{ix})\right) - \ln\left(\pi_{1/1} + \sum_{i=1}^{5} \pi_{i/1} \exp(\beta_{ix})\right) \quad (2)$$

for $j = 2, 3, \ldots, 5$ respectively.

In this equation, the expressions of the naive coefficients are explained by the corrected coefficients; however, our interest is the corrected coefficients explained by the naive coefficients. We thus had to solve a system of 5–1 equations with 5–1 unknowns, which is provided by the above expressions.

Using the fifths for socioeconomic categories, the system of Eq. (2) becomes as follows:

$$\begin{cases} RR_{2w} = \dfrac{\pi_{1/2} + \pi_{2/2}RR_{2x} + \pi_{3/2}RR_{3x} + \pi_{4/2}RR_{4x} + \pi_{5/2}RR_{5x}}{\pi_{1/1} + \pi_{2/1}RR_{2x} + \pi_{3/1}RR_{3x} + \pi_{4/1}RR_{4x} + \pi_{5/1}RR_{5x}} \\ RR_{3w} = \dfrac{\pi_{1/3} + \pi_{2/3}RR_{2x} + \pi_{3/3}RR_{3x} + \pi_{4/3}RR_{4x} + \pi_{5/3}RR_{5x}}{\pi_{1/1} + \pi_{2/1}RR_{2x} + \pi_{3/1}RR_{3x} + \pi_{4/1}RR_{4x} + \pi_{5/1}RR_{5x}} \\ RR_{4w} = \dfrac{\pi_{1/4} + \pi_{2/4}RR_{2x} + \pi_{3/4}RR_{3x} + \pi_{4/4}RR_{4x} + \pi_{5/4}RR_{5x}}{\pi_{1/1} + \pi_{2/1}RR_{2x} + \pi_{3/1}RR_{3x} + \pi_{4/1}RR_{4x} + \pi_{5/1}RR_{5x}} \\ RR_{5w} = \dfrac{\pi_{1/5} + \pi_{2/5}RR_{2x} + \pi_{3/5}RR_{3x} + \pi_{4/5}RR_{4x} + \pi_{5/5}RR_{5x}}{\pi_{1/1} + \pi_{2/1}RR_{2x} + \pi_{3/1}RR_{3x} + \pi_{4/1}RR_{4x} + \pi_{5/1}RR_{5x}} \end{cases} \quad (3)$$

where $RR_{jw} = \exp(\beta_{jw})$ and $RR_{ix} = \exp(\beta_{ix})$ represent the relative risks. After resolution of the system, we obtained the corrected relative risks in terms of the naive relative risks and misclassification coefficients. The expressions are as follows:

$$\begin{cases} RR_{2x} = \dfrac{a_1 + a_2RR_{2w} + a_3RR_{3w} + a_4RR_{4w} + a_5RR_{5w}}{m_1 + m_2RR_{2w} + m_3RR_{3w} + m_4RR_{4w} + m_5RR_{5w}} \\ RR_{3x} = \dfrac{b_1 + b_2RR_{2w} + b_3RR_{3w} + b_4RR_{4w} + b_5RR_{5w}}{m_1 + m_2RR_{2w} + m_3RR_{3w} + m_4RR_{4w} + m_5RR_{5w}} \\ RR_{4x} = \dfrac{c_1 + c_2RR_{2w} + c_3RR_{3w} + c_4RR_{4w} + c_5RR_{5w}}{m_1 + m_2RR_{2w} + m_3RR_{3w} + m_4RR_{4w} + m_5RR_{5w}} \\ RR_{5x} = \dfrac{d_1 + d_2RR_{2w} + d_3RR_{3w} + d_4RR_{4w} + d_5RR_{5w}}{m_1 + m_2RR_{2w} + m_3RR_{3w} + m_4RR_{4w} + m_5RR_{5w}} \end{cases} \quad (4)$$

The values of the coefficients a_1, a_2, a_3, a_4, a_5, b_1, b_2, b_3, b_4, b_5, c_1, c_2, c_3, c_4, c_5, d_1, d_2, d_3, d_4, d_5, m_1, m_2, m_3, m_4, m_5 are given in the Appendix. The system was solved using the Maple software. The extension of Eq. (4) to numbers of categories other than five is straightforward.

3. Application using French data

3.1. Evaluation of misclassification in France

To assess misclassification due to changing address (the $\pi_{i/j}$ in our notations), we used a database extracted from the Institut National de la Statistique et des Etudes Economiques (INSEE) National Census 1990 and 1999. The database, which is distributed by the Centre Maurice Halbwachs, documents movements between the municipality of residence at the time of the 1990 census and that of the 1999 census for the whole of France. More precisely, for each place of residence in 1999, the data indicates the proportion of the population living in the same municipality in 1990, and details of the origins of the inhabitants from another municipality. The social deprivation of the various municipalities was assessed using the quintiles of the French version of the European Deprivation Index [15]. With respect to the