Breast cancer incidence using administrative data: correction with sensitivity and specificity
Chantal Marie Courisa,d,*, Stephanie Polazziia, Frederic Olivelb, Laurent Remontetc,d,e, Nadine Bossardc,d,e, Frederic Gomezf, Anne-Marie Schotta,d, Nicolas Mittong, Marc Colonna, Beatrice Tromberti
aPole Information Médicale Evaluation Recherche, Hôpitaux Civils de Lyon, Université de Lyon, Santé Individu Société, Lyon, France
bDépartement d’Information Médicale, CHU de Grenoble, Grenoble, France
cHôpitaux Civils de Lyon, Service de Biostatistique, Lyon, France
dUniversité de Lyon, Université Lyon I, Villeurbanne, France
eUMR CNRS 5558, Laboratoire Biostatistique Santé, Pierre-Bénite, France
fDépartement d’Information Médicale, Centre Léon Bérard, Lyon, France
gRegistre des cancers de l’Isère, France
hFRANCIM, Toulouse, France
iService de Santé Publique et d’Information Médicale, CHU de Saint-Etienne, France

Accepted 28 July 2008

Abstract

Objective: To estimate breast cancer incidence in the general population using a method that corrects for lack of sensitivity and specificity in the identification of incident breast cancer in inpatient claims data.

Study Design and Settings: Two-phase study: phase 1 to identify incident cases in claims data, and phase 2 to estimate sensitivity and specificity in a subset of the population. Two algorithms (1: principal diagnosis; 2: principal diagnosis + specific surgery procedures) were used to identify incident cases in claims of women aged 20 years or older, living in a French district covered by a cancer registry. Sensitivity and specificity were estimated in one district and used to correct incident cases identified.

Results: The sensitivity and specificity for algorithms 1 and 2 were 69.0% and 99.89%, and 64.4% and 99.93%, respectively. In contrast to specificity, the sensitivity for both algorithms was lower for women younger than 40 years and older than 65 years. Cases reported by cancer registries were closer to cases identified with algorithm 2 (−3.2% to +20.1%) and to corrected numbers with algorithm 1 (−1% to +15%).

Conclusion: To obtain reliable estimates of breast cancer incidence in the general population, sensitivity and specificity, which reflect medical and coding practice variations, are necessary. © 2009 Elsevier Inc. All rights reserved.

Keywords: Measurement error; Claims databases; Biostatistics; Cancer; Incidence; Epidemiology

1. Introduction

Data collected by population-based cancer registries are usually considered the “gold standard” and are used to estimate cancer incidence throughout the world [1,2]. However, many countries, such as the United States of America, the United Kingdom, Ireland, Switzerland, Spain, Italy, or France do not have a nationwide population-based cancer registry. In France, there are nine general and seven site-specific population-based cancer registries. They describe cancer incidence for a 15% nonrandom sample of the French population and provide modeled national incidence estimates [3]. Although these estimates can be used to plan cancer prevention and intervention programs at the national level, the evaluation of these action programs and the monitoring of cancer incidence require information on the population for which the program is directed.

In most countries, claims databases collect data on all stays of a patient in a healthcare facility using standardized discharge abstracts that include mandatory information on the patients, their diseases, and the procedures performed [4]. The French equivalent of claims data is obtained from the “Programme de Médicalisation des Systèmes d’Information,” and could provide national estimates of cancer
incidence. The fact that claims databases were originally intended for reimbursement rather than to provide information on patients’ health status might introduce bias in the identification of incident cancer cases. This bias might be related to false incident cancers because of a lack of specificity, and to unrecorded cancers, leading to a lack of sensitivity [5–7].

A large number of studies evaluated the reliability of hospitals’ claims databases in identifying incident cancers using algorithms based on diagnosis codes and on diagnosis and procedural codes [8–33]. A limitation was the ability to identify incident cases. This was related to miscoding of medical information, such as incorrect case identification owing to incorrect diagnosis codes of breast cancer from a negative biopsy [10]. However, the main challenge has been to differentiate incident from prevalent cases [8–12,15–17,25,26,33].

Because of all the known imperfections of claims data to identify incident cancer cases, a true or exact measurement of the number of incident cancer cases cannot be made [34,35]. To correct this measurement, we used a statistical method based on a two-phase study design that we had previously developed [36,37]. The first phase aimed to identify incident cases of breast cancer in the studied population using inpatient claims data, and the second phase was designed to validate the identification in a subset of the population using cancer registry data. The validation involved estimating the sensitivity and specificity of identifying incident breast cancer based on claims data using a cancer registry as the gold standard [38]. Our goal was to validate this statistical method of estimating breast cancer incidence in a general population. The method was applied to the French population covered by a cancer registry to compare the number of cases obtained by both sources.

2. Methods

2.1. Studied population

The analysis included all women 20 years of age or older living in one of the nine districts covered by a cancer registry in 2002. Metropolitan France is administratively divided into 96 districts (“Départements”), identified by a name and a number, and most of the French cancer registries have a geographic area determined by districts. In 2002, there were nine districts covered by a general cancer registry. They provided information on cancer incidence for a 13% nonrandom sample of the French population [27]. The districts covered were 25-Doubs, 34-Herault, 38-Isère, 44-Loire-Atlantique, 50-Manche, 67-Bas-Rhin, 80-Somme, 81-Tarn, and 85-Vendee.

2.2. Data set

Data were obtained from the French National Institute of Statistics and Economic Studies [39], the French equivalent of the Diagnosis Related Group programme, the “Programme de Médicalisation des Systèmes d’Information” (PMSI) and the French cancer registries association (FRANCIM).

The French National Institute of Statistics and Economic Studies provides annual population figures on January 1 of women aged 20 years or older in France by district.

The PMSI-provided database includes standard inpatient discharge abstracts containing compulsory information on patients (gender, age, and postal codes), the diagnoses (using ICD-10 codes) and the care received (using procedural codes of the French Medical Acts Catalogue) [40] for each patient stay in any health care center in France. The principal diagnosis is defined as the diagnosis that contributed the most to the care provided during hospitalization. A unique anonymous patient identification number is attributed to each inpatient.

FRANCIM, in collaboration with the Biostatistics unit of the Hospices Civils de Lyon and the Institut de Veille Sanitaire, has a common database of all cases diagnosed and registered in all French districts covered by a cancer registry.

The incident cases in the studied population were identified in the 2002 claims data. The identified and corrected incident cases were then compared with the 2002 registry data. The validation database was built with data from the 38-Isère registry and claims data for the year 2001. This validation database allowed the estimation of sensitivity and specificity (see in the following section).

2.3. Identification of incident cancer cases in claims data

Two algorithms were designed to identify incident breast cancer cases in the hospital stays of women aged 20 years or older living in one of the nine districts covered by a cancer registry in 2002. The first algorithm selected women having a hospital stay with a principal diagnosis code of invasive breast cancer (ICD-10 codes C50.0–C50.9). The second algorithm selected women having a hospital stay with a principal diagnosis code of breast cancer and procedural codes for breast cancer–specific surgery. The number of cases identified in the PMSI-provided database for each district was compared with that reported by the corresponding cancer registry. To quantify the relative difference for each district, the difference between the number of cases in the registry and the number identified in the PMSI-provided data was divided by the number of cases in the registry.

2.4. Validation database

Personal identifiers (last name, first name, date of birth) of women living in the 38-Isère district and having a stay with diagnosis codes of breast cancer were obtained from the healthcare center where they had been hospitalized in
2001 and entered into the validation database. These personal identifiers and those collected by the cancer registry of the 38-Isère district were anonymized using the ANONYMAT software developed by the Department of Medical Informatics of the teaching hospital of Dijon [41,42]. Both anonymous personal identifiers were then compared using a probabilistic matching [41,42] adapted from the “AUTO-MATCH” method [43].

2.4.1. Sensitivity and specificity

The cancer registry of the 38-Isère district was considered as the gold standard to identify incident cancer cases. Using the validation data set, the sensitivity (Se) was defined as the percentage of breast cancer cases in the cancer registry of the Isère district that had been identified from the PMSI-provided databases. Specificity (Sp) was the percentage of women who had been identified as not having incident breast cancer from the cancer registry of the Isère district, and who did not have a breast cancer from the PMSI-provided databases. Se and Sp were estimated for both algorithms for several age groups; these estimates have been smoothed using splines to obtain more reliable estimates [44].

2.5. Corrected number of incident breast cancer cases

A corrected number of incident breast cancer cases (~K) was obtained using the following formula:

\[ \sim K = \frac{K_{\text{rec}} - (1 - Sp)N}{Se + Sp - 1} \]

where \( N \) is the number of women aged 20 years or older, living in the general population in 2001; \( K_{\text{rec}} \) is the number of women identified in the PMSI-provided database with one of the algorithms among \( N \). The statistical method was developed and validated by simulation as in a previous study [38].

2.5.1. Validation of the corrected number

Corrected numbers were estimated by age group using both algorithms to identify cases in the PMSI-provided database for each French district covered by a general population-based cancer registry. A prediction error (PE) was calculated for each district using the following formula:

\[ PE = \sum_{i=1}^{n} \left( \frac{K_i - \hat{K}_i}{\sqrt{\hat{K}_i}} \right)^2, \]

where \( i \) is the number of age group, \( K \) the number of cases reported by the cancer registries and \( \hat{K} \) the corrected number. PE follows a chi-squared distribution with \( n \) degrees of freedom. For each district, the corrected number was compared with the number reported by the corresponding cancer registry. A relative difference was quantified as previously described.

3. Results

3.1. Study population

For the 2,590,395 women aged 20 years or older, living in the nine French districts covered by a cancer registry, the first (principal diagnosis of invasive breast cancer) and the second (principal diagnosis of invasive breast cancer and at least one procedural code for breast cancer—specific surgery) algorithm identified 6,239 and 5,181 incident breast cancer cases, respectively, in the PMSI-provided database. The cancer registries for these districts reported 5,057 incident cases. Regarding each district, the number of incident cancer cases identified in the claims data with the first algorithm was consistently higher (minimum: +12%, maximum: +27%) than the number of cases reported by the cancer registries (Table 1). This systematic difference was not observed with the second algorithm. There was an overestimation among five districts (minimum: +1%, maximum: +15%) and an underestimation among four districts (minimum: −1%, maximum: −12%).

3.2. Validation database

For the 420,070 women aged 20 years or older living in the Isère district in 2001, 8,885 inpatient stays were identified in the claims database with a diagnosis code of breast cancer. The personal identifiers of these women were obtained for 8,247 (92.8%) hospital stays. After exclusion of the hospital stays without a personal identifier and selection of the first hospital stay for women, 995 women had had a stay with a principal diagnosis of invasive breast cancer (algorithm 1), and 816 had had a stay with both a principal diagnosis code of invasive breast cancer and at least one procedural code for breast cancer—specific surgery (algorithm 2). Out of the 825 women with an incident breast cancer reported by the cancer registry of the Isère district, 569 (69%) and 529 (64.1%) were, respectively, identified in the claims data by the first and the second algorithm (Table 2).

The number of false negatives was 15% lower with the first algorithm. Regarding false negatives of both algorithms, 172 women (algorithm 1: 67.1%; algorithm 2: 58.1%) were not found in the PMSI-provided database. Of these women, 84 (48.8%) had their cancer registered as incident by the Isère cancer registry in September to December 2001.

The number of false positives was more than 30% higher with the first algorithm, although the specificity appeared similar for both algorithms. The number of false positives with a date of discharge in the first 3 months of 2001 was 253 (59.4%) with the first algorithm and 163 (56.8%) with the second algorithm. Personal identifiers were missing for 138 (32.4%) and 126 (43.9%) of the false positives with the first and the second algorithm, respectively.
For both algorithms, sensitivity and specificity of identification of new breast cancer cases in the claims data varied with age (Fig. 1). The variations were higher for sensitivity (35.7% to 97.8%) than for specificity (99.73% to 99.99%). For both algorithms, the lowest sensitivities were observed for women younger than 40 years and older than 65 years. The specificities of both algorithms decreased for women aged 40 years or older and were the lowest for women aged 65 years or older.

### 3.3. Corrected number of incident breast cancer cases

The number of corrected cases was significantly different from that reported by the French cancer registries, as the PE exceeded the limit value for most of the district. The first algorithm tended globally to either over or underestimate the number of cases depending on the district observed, whereas the second algorithm mainly overestimated (−11.5% to +20.1%) (Table 1).

<table>
<thead>
<tr>
<th>District</th>
<th>Registry</th>
<th>Algorithm 1&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Algorithm 2&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>825</td>
<td>995</td>
<td>821</td>
</tr>
<tr>
<td>25</td>
<td>356</td>
<td>444</td>
<td>355</td>
</tr>
<tr>
<td>34</td>
<td>734</td>
<td>924</td>
<td>770</td>
</tr>
<tr>
<td>44</td>
<td>941</td>
<td>1,161</td>
<td>1,009</td>
</tr>
<tr>
<td>50</td>
<td>316</td>
<td>389</td>
<td>260</td>
</tr>
<tr>
<td>67</td>
<td>788</td>
<td>1,006</td>
<td>879</td>
</tr>
<tr>
<td>80</td>
<td>402</td>
<td>508</td>
<td>415</td>
</tr>
<tr>
<td>81</td>
<td>262</td>
<td>327</td>
<td>236</td>
</tr>
<tr>
<td>85</td>
<td>433</td>
<td>485</td>
<td>346</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5,057</td>
<td>6,239</td>
</tr>
</tbody>
</table>

Abbreviations: PE, prediction error; PMSI, Programme de Médicalisation des Systèmes d’Information.

<sup>a</sup> Cases identified from the PMSI-provided data with breast cancer as the principal diagnosis.

<sup>b</sup> Cases identified from the PMSI-provided data with breast cancer as the principal diagnosis and procedural codes for breast cancer—specific surgery.

<sup>c</sup> PE follows a chi-squared distribution, and the threshold value to reach 5% statistical significance is 21.03.

### 4. Discussion

In this study, two algorithms were tested. The first one using principal diagnosis of breast cancer was more sensitive but less specific than the second one, which considered principal diagnosis of breast cancer and procedural codes for breast cancer—specific surgery. With both algorithms, the main difficulty was distinguishing incident cases from prevalent ones, because a wide majority of false positives and false negatives were, respectively, retrieved in the first and last 3 months of the study period. The cases found in the first 3 months might have been in the cancer registry in the former year, whereas the cases found in the last 3 months might already be in the registry but not yet hospitalized for their cancer or discharged from their hospital stay for their cancer. We were not able to confirm this hypothesis for several reasons, such as data availability and linkage feasibility, but this was confirmed by Warren et al. [11]. This would not have modified our sensitivity and specificity estimates. Even if our results cannot easily be compared with other published studies because of differences in the claims data used, the study period, and cancer stage, our results are similar to those reported by several other studies [8–11].

This study reported that the sensitivity and specificity to identify incident breast cancer cases in claims data varied with age. The ability of claims data to identify incident breast cancer cases is lower for women younger than 45 years or older than 65 years. This result was found regardless of the algorithm used. Similar findings have been observed in other studies, which aimed to evaluate Medicare and Medicaid data [18,27,29,45]. There are a number of hypotheses that could explain these variations with age. The
first is directly related to the variation of breast cancer incidence by age group. The incidence is lower for women younger than 45 years or older than 75 years. Thus, in these age groups, small variations in the identification of incident breast cancer in claims data may have a higher impact on sensitivity and specificity estimates than in the other age groups. The second hypothesis is that these variations might be explained by the differences in treatment patterns owing to age. For the oldest women, the lower sensitivity and the higher specificity may be related to less extensive or aggressive treatment owing to the frailty of such patients or to the existence of comorbid illness. For the youngest women, it can be attributed to the existence of several other breast diseases, which can be coded instead of cancer. In two other studies [29,45], sensitivity and specificity were also related to ethnicity, cancer stage, study period, and area. Regarding the study period, sensitivity and specificity were estimated for the year 2001, whereas the incident breast cancer cases were identified in the claims data in 2002. We cannot exclude the possibility that sensitivity and specificity may vary over time. For example, the French health care centers progressively adopted modifications implemented in 2001. Since 2004, medical activity measures in French claims data are directly related to funding of health care centers. This direct relationship to payment may be associated with increasing quality control that might homogenize and improve the data quality over time and in different areas. Regarding the area, the present study showed that the sensitivity and specificity estimates obtained in the Isère district were not systematically valid to estimate the number of incident breast cancer cases in the other districts covered by a cancer registry. The non-applicability of the sensitivity and specificity obtained in the Isère district to the other district might be related to variations in medical and coding practices between districts. Even though the French cancer registries tend to standardize their collection procedures by following the guidelines edited by the International Agency for Research on Cancer, there may be some variations between registries that can influence our result. To obtain a better assessment of these variations and to estimate breast cancer incidence in districts not covered by a cancer registry, sensitivity and specificity should be estimated using all French cancer registries as a gold standard. These estimates are difficult to obtain in France because both claims and registry data are anonymous, and the keys used to anonymize the data are different and may not directly be used for linkage. Thus, it is necessary to obtain personal identifiers, a challenge for routine analyses attributed to the need to visit each health care center and then to perform primary data collection. This difficulty is even more complicated, because the health care centers have different systems to maintain their claims data and do not have professionals to provide the requested information. In our study, the sensitivity and specificity might be biased, because we did not obtain all the personal identifiers to link the Isère cancer

![Algorithm 1: cases identified from the PMSI-provided data with breast cancer as the principal diagnosis sensitivity](image1)

![Algorithm 2: cases identified from the PMSI-provided data with breast cancer as the principal diagnosis and procedural codes for breast cancer specific-surgery sensitivity](image2)

**Fig. 1.** Sensitivity and specificity by 5-year age groups of women identified from the 2001 Programme de Médicalisation des Systèmes d’Information (PMSI)-provided database as having incident breast cancer matched with breast cancer cases reported by the Isère cancer registry.
registry to the claims data. Even if the percentage of missing data is rather low, most of the cases were for private hospitals. However, the method provided in this article can be valid for countries that have a valid linkage of their claims data to cancer registries (i.e., SEER-Medicare).

When the number of cases identified by both algorithms was compared with the number of cases reported by the cancer registries, the first algorithm systematically overestimated the number of cases. The second algorithm was less systematic, but the number of cases identified was closer to the number of cases reported by the cancer registries. Because the bias with the first algorithm is always in the same direction, this algorithm might be the best one to apply the correction method, whereas the second algorithm provides a better estimation of the number of incident cases without correction. To use the correction method to estimate cancer incidence in a general population, sensitivity and specificity that reflect a representative range of medical coding and registration practice variations in that population are necessary.

Acknowledgments

This study was supported by grants from the Cancérople Lyon Auvergne Rhône-Alpes (CLARA), the Centre International de Recherche en Economie de la Santé (CIRIES) and from Institut National du Cancer (INCA). The authors wish to thank the participating health care centers: Clinique Saint Jean (Dr. André), Clinique Jeanne d’Arc (Dr. André), Clinique Saint Charles (Dr. Perrot), Clinique Saint Vincent de Paul (Dr. Carriot), Polyclinique Pasteur (Dr. Bolzinger), Clinique Belledonne (Dr. Vial), Institut Privé de Cancérologie (Dr. Jacquin and Mrs. Tramanglia), Centre Hospitalier de Chambéry (Dr. Mercier), Clinique Mutualiste des Eaux Claires (Dr. Elsener and Mr. Macutan), Centre Hospitalier de Voinon (Dr. Ghiotti), Centre Hospitalier de Saint Marcellin (Dr. Bonis and Mr. Bonnay), Clinique la Parisière (Dr. Carriot), Clinique Charcot (Dr. Datra), Clinique du grand large (Dr. Cariot and Mrs. Angelot), Clinique Trarieux (Dr. Gillet), Clinique Trezel (Dr. Perrot), CH Bourgoin (Dr. Ducarre), Clinique mutualiste Eugène André (Mrs. Guiraud), Centre Hospitalier de Vienne (Dr. Lapierre). We are grateful to the Department of Biostatistics and Medical Informatics of the teaching hospitals of Dijon (Dr. Quantin and F. Benzeno) for the use of ANONYMAT and to Dr. S. Jaglal and S. Munce for their very helpful review of the article.

References


http://www.atih.sante.fr/?id=0003200003FF.


