The Accuracy of Japanese Claims Data in Identifying Breast Cancer Cases

Izumi Sato,* a,† Hiroshi Yagata, b and Yasuo Ohashi*‡

a Epidemiology and Statistics, Graduate School of Medicine, The University of Tokyo; 7–3–1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan; and b Breast Surgical Oncology, St. Luke’s International Hospital; 9–1 Akashi-cho, Chuo-ku, Tokyo 104–8560, Japan.

Received July 29, 2014; accepted November 3, 2014

In use of a claims database for a study, an inaccurate diagnosis of breast cancer based on claims data may lead to invalid study results. The aim of this study was to assess the accuracy of definitions for identifying breast cancer cases from the Japanese claims database. The study cohort consisted of women with no prior cancer-related history, from the claims data at a single institution between January 1 and December 31, 2011. We developed 14 definitions for identifying breast cancer based on claims data, using a combination of diagnosis codes and treatment procedure codes. We calculated the sensitivity, specificity, and positive predictive value (PPV) of each definition, compared to cases identified from the standardized hospital-based cancer registry as a standard reference. A total of 50056 women were included in the study cohort from the claims database. We identified 633 breast cancer cases from the cancer registry. Of 14 definitions, 12 exhibited higher sensitivity than 90%, while the others exhibited lower sensitivity than 40%. The specificities of all definitions were high (≥99%), and the PPVs were between 65.8 and 90.7%. We selected the most optimal definition obtained from combinations of diagnosis and cancer treatment codes (surgery, chemotherapy, medication, radiation procedure), which had high values for sensitivity (90.4%), specificity (99.8%), and PPV (87.3%). Definitions obtained via combinations of the diagnosis codes and procedure codes could be used to accurately identify breast cancer cases from the claims database. Further studies in a multi-institutional setting are planned to confirm our results.

Key words claims data; diagnosis; breast cancer; cancer registry

Large electronic databases, including claims databases and medical record databases have been used for pharmacoepidemiologic studies (e.g., evaluation of drug utilizations, research for adverse drug reactions, policy evaluations). In Japan, both the utilization of healthcare utilization databases and infrastructure development lag behind those of other countries. Recently, a National Database (NDB) containing claims data from throughout Japan has been constructed.2) This NDB includes nearly all Japanese claims, as Japan adopts a universal health care system.3) Japanese claims data are recorded electronically using computer processing master codes that are unique to Japan,4) rather than international standardized codes (e.g., the International Classification of Diseases (ICD)).5) International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)).6) Although use of the database for various studies is expected to grow, this use is currently limited by the presence of stringent criteria (e.g., prohibition against linking with other sources).7)

Secondary use of claims data presents several strengths. This data contains substantial healthcare information which is free from recall and interview bias, and allows researchers to obtain large data collection sample sizes at low cost.1,8) In contrast, such use also presents weaknesses, including lack of data such as potential confounders, examination results, and disease severity, as well as uncertain diagnoses, as claims are made for the purpose of fee-for-service reimbursement, rather than for research.9) To obtain additional information, a claims database is sometimes linked with another databases, however, databases are not always linkable. Researchers should assess the completeness and accuracy of the key study variables such as diagnosis, in terms of whether the quality of the variables is adequate for the completion and interpretation of a particular analysis.9) We are planning to evaluate the breast cancer preventive benefits of a medication using a large claims database. Thus, it is necessary to assess the accuracy of breast cancer diagnoses based on claims data, in order to adopt it as the study outcome.

Previous studies have attempted to evaluate the accuracy of breast cancer diagnoses on claims data.10–19) For example, Freeman et al.17) determined the optimal combination of predictors (e.g., diagnosis, procedures) using a logistic regression model to identify the incidence of breast cancer from a population-based cancer registry and Medicare claims data, which included data of inpatients, outpatients, and physician services. Setoguchi et al.18) developed case definitions to identify 6 types of cancer cases using multiple combinations of predictors in Pennsylvania, using Medicare claims data linked to pharmacy dispensing data and the population-based cancer registry. Couris et al.19) found that an algorithm consisting of diagnosis and breast cancer-related surgery codes could identify a similar number of the cases reported by the cancer registry.

Assessing the quality of diagnosis is also necessary for the Japanese claims database. It is difficult to apply methods from previous studies to the Japanese system, due to the presence of many differences in the medical environment (e.g., insurance system, policy, race, and mortality) and recording styles.
of claims (e.g., variables, codes) between Japan and other countries. Therefore, we developed definitions to identify breast cancer cases from the claims database and assessed the accuracy of these definitions.

**MATERIALS AND METHODS**

**Data Sources** Data from the claims database and the standardized hospital-based cancer registry was obtained from St. Luke’s International Hospital, a designated cancer care hospital in Japan.20) We identified a cohort of subjects who have at least one claim between January 1 and December 31, 2011.

Since approval of the Cancer Control Act in 2006, designated cancer care hospitals have been developed for providing standardized high-quality cancer care in each prefecture (Japan is comprised of 47 prefectures).21) These hospitals are required to meet the following requirements in Japan: improving the quality of cancer care; playing a central role in the development of a cooperative system in each prefecture; providing consultation support and information for cancer patients; and, promoting the standardized hospital-based cancer registry.20)

This hospital also meets the requirements for advanced treatment hospitals, and has adopted the Diagnosis Procedure Combination (DPC) comprehensive medical fee evaluation system,22,23) similar to the Diagnosis Related Group/Prospective Payment system (DRG/PPS) of Medicare.24) A single medical fee for each DRG is calculated based on a patient’s discharge diagnosis. Although the medical fee of the DPC system is also determined via comprehensive assessment, a fee for each patient is calculated based on the combination of diseases for which the greatest amounts of resources were spent, during the hospitalization and procedures.22,23)

The claims data are recorded electronically in CSV format, using a set of computerized processing master codes which are original to Japan, and are based on a specialized method established by the Ministry of Health, Labour and Welfare of Japan.25) Claims include information on patient characteristics (gender, birth date, type of insurance), diagnosis (Japanese original master codes, DPC-based codes (for hospitalized patients only)), ICD10th Revision (ICD-10) codes (for hospitalized patients only)), medical treatments procedures (Japanese original master codes), prescriptions (Japanese original master codes), and hospital admissions (periods of hospitalization).26)

This study was approved by the Institutional Review Board of St. Luke’s International Hospital.

**Cases from the Cancer Registry** We defined breast cancer cases with an ICD for Oncology, third edition (ICD-O-325)) topography code of breast cancer (C500 to C506, C508, C509) registered on the standardized hospital-based cancer registry between January 1 and December 31, 2011. We used the cases as a standard reference except three cases without claims data. We used the standardized hospital-based cancer registry because all designed cancer care hospitals register cancer cases using a standard set of procedures and registration form26,27) and the hospitals use various information to enhance the accuracy of finding cancer cases (e.g., pathologic reports, discharge summaries, chemotherapy records, surgery records). A target tumor for the registry corresponds to a behavioral code of 2 or 3 in the ICD-O-3.

**Case Definitions for the Claims Data** Because we had no information on the Japanese codes used in the claims of breast cancer patients, prior to constructing definitions, we investigated the range of codes used in the claims data of breast cancer patients. We preliminarily investigated one hundred patients (arbitrarily determined number) with breast cancer, randomly selected from the cancer registry to explore the code profile, as a reference to develop definitions to identify breast cancer. This patient number was considered sufficient for a rough assessment of the profile. The codes used in these claims were classified into nine code groups: diagnosis (DPC codes, Japanese original codes), breast cancer-related diagnosis (e.g., cancer pain), surgery, marker test, chemotherapy, biopsy, imaging, medication, radiation therapy, other breast cancer-related codes which were selected in advance by a review of previous studies (Appendix). Some codes (Japanese original codes; counseling or therapeutic management for any cancer) were also added which were considered to be related to breast cancer.

We exploratorily developed 14 definitions to identify breast cancer cases from the claims data using nine code groups (Table 1). Definitions 1 and 14 are based on only diagnosis codes, Definitions 2 to 10 are based on diagnosis codes

<table>
<thead>
<tr>
<th>Definition</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Diagnosis of breast cancer</td>
<td></td>
</tr>
<tr>
<td>2 Diagnosis of breast cancer + diagnosis code</td>
<td>related to breast cancer</td>
</tr>
<tr>
<td>3 Diagnosis of breast cancer + diagnostic imaging code</td>
<td></td>
</tr>
<tr>
<td>4 Diagnosis of breast cancer + biopsy code</td>
<td></td>
</tr>
<tr>
<td>5 Diagnosis of breast cancer + marker test code</td>
<td></td>
</tr>
<tr>
<td>6 Diagnosis of breast cancer + surgery code</td>
<td></td>
</tr>
<tr>
<td>7 Diagnosis of breast cancer + chemotherapy code</td>
<td></td>
</tr>
<tr>
<td>8 Diagnosis of breast cancer + medication code</td>
<td></td>
</tr>
<tr>
<td>9 Diagnosis of breast cancer + radiation procedure code</td>
<td></td>
</tr>
<tr>
<td>10 Diagnosis of breast cancer + the other code related to breast cancer</td>
<td></td>
</tr>
<tr>
<td>11 Diagnosis of breast cancer + diagnosis code related to breast cancer or marker test code</td>
<td></td>
</tr>
<tr>
<td>12 Diagnosis of breast cancer + surgery, chemotherapy, medication, or radiation procedure code</td>
<td></td>
</tr>
<tr>
<td>13 Diagnosis of breast cancer + diagnosis code related to the breast cancer, marker test code, surgery, chemotherapy, medication or radiation procedure code</td>
<td></td>
</tr>
<tr>
<td>14 ≥3 Diagnoses of breast cancer</td>
<td></td>
</tr>
</tbody>
</table>

---

20) St. Luke’s International Hospital, a designated cancer care hospital in Japan.
21) These hospitals are nated cancer care hospitals have been developed for providing cancer care in Japan.
22) Similar to the Diagnosis Related Group/Prospective Payment system (DRG/PPS) of Medicare.
23) A single medical fee for each DRG is calculated based on a patient’s discharge diagnosis.
24) The medical fee of the DPC system is also determined via comprehensive assessment, a fee for each patient is calculated based on the combination of diseases for which the greatest amounts of resources were spent, during the hospitalization and procedures.
25) Diagnosis codes, Definitions 1 and 14 are based on only diagnosis codes, Definitions 2 to 10 are based on diagnosis codes, and Definitions 11 and 12 are based on diagnosis codes related to breast cancer.
26) The codes used in these claims were classified into nine code groups: diagnosis (DPC codes, Japanese original codes), breast cancer-related diagnosis (e.g., cancer pain), surgery, marker test, chemotherapy, biopsy, imaging, medication, radiation therapy, other breast cancer-related codes which were selected in advance by a review of previous studies (Appendix). Some codes (Japanese original codes; counseling or therapeutic management for any cancer) were also added which were considered to be related to breast cancer.

---
and the types of procedure codes, and Definitions 11 to 13 are based on combinations of diagnosis codes and multiple types of procedure codes. Regarding the diagnosis codes for breast cancer, we used 2 different types of codes (DPC codes and Japanese original codes). Most inpatient diagnoses were recorded in DPC codes, while outpatient diagnoses were recorded in Japanese original codes. Additionally, we considered “primary flag” records as diagnosis codes. A “primary flag” is arbitrarily recorded before the assignment of a definitive diagnosis, then unrecorded once a definite diagnosis has been provided. To exclude suspicious breast cancer cases, we eliminated cases with a “primary flag.”

**Data Linkage** Anonymized patient ID numbers were only used to protect patient privacy and to establish a link between the claims data and the standardized hospital-based cancer registry. All breast cancer cases from the cancer registry could be linked to the claims data.

**Analysis** We calculated the sensitivity (1), specificity (2), positive predictive value (PPV) (3), and 95% confidence intervals (CIs) for each definition. A true positive (TP) was defined as a case that was identified by the case definition and also identified in the standardized hospital-based cancer registry, a false negative (FN) as a case that was not identified in the standardized hospital-based cancer registry but not by the case definition, and a true negative (TN) as a case that was not identified by the case definition, and was also not identified in the standardized hospital-based cancer registry. Sensitivity is the ratio of TP to the sum of TP and FN (1), Specificity is the ratio of TN to the sum of TN and FP (2), and PPV is the ratio of TP to the sum of TP and FP (3). In addition, we made exploratory speculations regarding the optimal combination of these quantities.

\[
\begin{align*}
\text{Sensitivity} &= \frac{TP}{TP + FN} \\
\text{Specificity} &= \frac{TN}{TN + FP} \\
\text{PPV} &= \frac{TP}{TP + FP}
\end{align*}
\]

**RESULTS**

**Characteristics of the Study Cohort** We extracted 50056 subjects in the study cohort. We identified 633 breast cancer cases from the standardized hospital-based cancer registry to which we could link claims (Table 2). The mean (±SD) age of the cases was 51.0 (±11.8) years. A total of 630 of 633 cases (99.5%) were diagnosed as breast cancer confirmed by histology. The most common treatment was surgery (523 cases, 82.6%), followed by endocrine therapy (334 cases, 52.8%), radiation therapy (244 cases, 38.5%), and chemotherapy (182 cases, 28.8%).

**Accuracy of Definitions** Definitions 1, 3, 4, and 11–14 exhibited higher sensitivity than 90% (Table 3). Of these, Definitions 1, 3, 4, and 13 had accuracies higher than 95%. In contrast, Definitions 5 and 7–9 exhibited lower sensitivity than 70%. In particular, Definitions 7–9 had the three lowest sensitivities (<37%). The specificities of all of the Definitions were higher than 99%. Definitions 5–9 had the highest specificities (99.9%). Definitions 2 and 5–14 had higher PPV than 80%; values for Definitions 2, 5, 6, 9, and 12 were higher than 85%, while Definition 6 had the highest PPV (90.7%). On the other hand, the PPV of Definitions 1 and 3 were lower than 70%.

Definitions 5, 6, and 9 had the three highest PPVs (90.7%, 87.8%, and 87.6%) but were in the lower half in terms of sensitivity (73.6%, 65.9%, and 32.4%), among all definitions. On the other hand, Definitions 1, 3, 4, and 13 had the four highest sensitivities (98.7%, 97.6%, 95.7%, and 97.3%) but were in the lower half in terms of PPV (65.8%, 69.8%, 78.6%, and 82.6%), among all definitions. Definitions which were in the upper half of all definitions for both sensitivity and PPV included Definition 11 (sensitivity: 94.2%, PPV: 84.5%) and Definition 12 (90.4%, 87.3%). In this regard, Definition 12 had the fourth highest PPV and was also in the upper half regarding sensitivity.

**DISCUSSION**

We developed case definitions for the identification the breast cancer from claims data using diagnosis codes and procedure codes, and calculated the sensitivity, specificity, and PPV for each definition *via* a comparison with the standardized hospital-based cancer registry as a gold standard. We identified the definitions which were capable of identifying breast cancer cases from the claims database with high sensitivity, specificity, and PPV.

| Characteristics of Breast Cancer Patients Who Were Identified in the Standardized Hospital-Based Cancer Registry |
|-----------------------------------------------------|---------|
| Characteristic | n=633   |
| Age | 51.0±11.8 |
| Age group at diagnosis |  
| ≤39 year | 95 (15.0)  
| 40–49 year | 236 (37.3)  
| 50–59 year | 153 (24.2)  
| 60–69 year | 104 (16.4)  
| ≥70 year | 45 (7.1)  
| Diagnostic confirmation |  
| Positive histology | 630 (99.5)  
| Positive cytology | 1 (0.2)  
| Unknown | 2 (0.3)  
| Surgical treatment |  
| Performed | 523 (82.6)  
| None | 97 (15.3)  
| Unknown | 13 (2.1)  
| Radiation therapy |  
| Performed | 244 (38.5)  
| None | 365 (57.7)  
| Unknown | 24 (3.8)  
| Endocrine therapy |  
| Performed | 334 (52.8)  
| None | 270 (42.7)  
| Unknown | 29 (4.6)  
| Chemotherapy |  
| Performed | 182 (28.8)  
| None | 450 (71.1)  
| Unknown | 1 (0.2)  

Table 2.
Because we had no information on the Japanese codes used in the claims of breast cancer patients prior to the construction of definitions, we developed exploratory definitions by reference to previous studies and using advice from experts in the area. We then evaluated each definition and attempted to determine the optimal combination of three validity measures (sensitivity, specificity, and PPV). An emphasis was placed on PPV, because an estimation of the risk ratio would be lower for high sensitivity than low sensitivity. In the current study, those definitions with the three highest PPVs (90.7%, 87.8%, and 87.6%) were among the 14 definitions. On the other hand, the definition with the fourth highest PPV (87.3%) was also in the upper half for sensitivity (90.4%). Furthermore, this PPV value was nearly the same as those of the second and third highest ranking definitions. Accordingly, Definition 12 (sensitivity: 90.4%, PPV: 87.3%) was considered as the optimal definition.

In previous studies, researchers have evaluated the accuracy of cancer diagnoses in claims databases, Freeman et al. obtained a high sensitivity (90%) and specificity (99%), but a low PPV (70%). In their discussion, they attributed the high sensitivity to the use of three kinds of claims data (inpatients, outpatients, and physician services), while the low PPV was considered to be due to limitations in distinguishing recurrent and secondary cancers. Setoguchi et al. developed a case definition using various combinations of diagnoses and procedure codes in Pennsylvania, obtaining a sensitivity, specificity, and PPV for the definition for identifying breast cancer cases of 83%, 99%, and 75%, respectively.

Our study indicated that the sensitivity and specificity of the most optimal definition was 90% and 99%, respectively. We consider that the high sensitivity was attributed to the use of both inpatient and outpatient claims data. Because inpatient claims data and outpatient claims data are separated in Japan, we are required to combine the two databases for use as integrated information. Thus, if we had used only the outpatient database, we might have obtained high sensitivity but low PPV. On the other hand, if we had used only the inpatient database, we might have obtained low sensitivity but high PPV. In addition, we also regard the high sensitivity to reflect the high proportion of patients who received treatment (e.g., 84% of cases received surgery); in particular, if many patients received no treatment (e.g., refusal of care, change of hospital), a lower sensitivity might be obtained.

The high PPV might be the result of potential selection bias, due to the characteristics of the designated cancer care hospital, because patients in designated cancer care hospitals have received, breast cancer diagnoses, in advance. In Japan, patients who are given a definitive cancer diagnosis at general hospitals are generally referred to a designated cancer care hospital to receive high-quality cancer care and support.

We had difficulty in accessing the NDB for study, due to the stringent requirement of permission for usage, and used two data sources for a single hospital. It is also difficult to access the national cancer registries consisting of the standardized hospital-based cancer registries in Japan, because of restrictions to the public. Furthermore, the NDB is banned from connection to other sources in the guidelines for its provision. Instead of the NDB, we linked the claims database with the cancer registry, to compare the cases identified from the claims database with those identified from the cancer registry. In our study, both data sources were provided by the same hospital, which allowed us to use the patient ID numbers for linkage.

This study has several limitations. First, we did not consider the potential for coding errors and the presence of non-exhaustive breast cancer cases from the hospital cancer registry. However, the recording style has been standardized in Japan, and this database is managed by coding specialists at the Personnel Department in the designated cancer care hospital. Therefore, the quality of the data is high. Second, as a single institutional study, it is unclear whether our results are directly applicable for other hospitals. Thirdly, because we did not evaluate all combinations in terms of accuracy, definitions which are more optimal than the study definitions may exist. However, we constructed those conceivable definitions which might have high values for the three study measures, according to previous studies and advice from experts in this field. The construction of additional definitions will be required.

Table 3. Sensitivity, Specificity, and PPV of Definitions of Breast Cancer

<table>
<thead>
<tr>
<th>Definition</th>
<th>Cases (n)</th>
<th>SEN (%)</th>
<th>95% CI</th>
<th>SPE (%)</th>
<th>95% CI</th>
<th>PPV (%)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>950</td>
<td>98.7</td>
<td>97.9–99.6</td>
<td>99.3</td>
<td>99.3–99.4</td>
<td>65.8</td>
<td>62.8–68.8</td>
</tr>
<tr>
<td>2</td>
<td>657</td>
<td>88.6</td>
<td>86.2–91.1</td>
<td>99.8</td>
<td>99.8–99.8</td>
<td>85.4</td>
<td>82.7–88.1</td>
</tr>
<tr>
<td>3</td>
<td>886</td>
<td>97.6</td>
<td>96.4–98.8</td>
<td>99.5</td>
<td>99.4–99.5</td>
<td>69.8</td>
<td>66.7–72.8</td>
</tr>
<tr>
<td>4</td>
<td>771</td>
<td>95.7</td>
<td>94.2–97.3</td>
<td>99.7</td>
<td>99.6–99.7</td>
<td>78.6</td>
<td>75.7–81.5</td>
</tr>
<tr>
<td>5</td>
<td>475</td>
<td>65.9</td>
<td>62.2–69.6</td>
<td>99.9</td>
<td>99.9–99.9</td>
<td>87.8</td>
<td>84.8–90.7</td>
</tr>
<tr>
<td>6</td>
<td>514</td>
<td>73.6</td>
<td>70.2–77.1</td>
<td>99.9</td>
<td>99.9–99.9</td>
<td>90.7</td>
<td>88.1–93.2</td>
</tr>
<tr>
<td>7</td>
<td>284</td>
<td>37.4</td>
<td>33.7–41.2</td>
<td>99.9</td>
<td>99.9–99.9</td>
<td>83.5</td>
<td>79.1–87.8</td>
</tr>
<tr>
<td>8</td>
<td>281</td>
<td>37.3</td>
<td>33.5–41.0</td>
<td>99.9</td>
<td>99.9–99.9</td>
<td>84.0</td>
<td>79.7–88.3</td>
</tr>
<tr>
<td>9</td>
<td>234</td>
<td>32.4</td>
<td>28.7–36.0</td>
<td>99.9</td>
<td>99.9–100.0</td>
<td>87.6</td>
<td>83.4–91.8</td>
</tr>
<tr>
<td>10</td>
<td>699</td>
<td>88.5</td>
<td>86.0–91.0</td>
<td>99.7</td>
<td>99.7–99.8</td>
<td>80.1</td>
<td>77.2–83.1</td>
</tr>
<tr>
<td>11</td>
<td>705</td>
<td>94.2</td>
<td>92.3–96.0</td>
<td>99.8</td>
<td>99.7–99.8</td>
<td>84.5</td>
<td>81.9–87.2</td>
</tr>
<tr>
<td>12</td>
<td>655</td>
<td>90.4</td>
<td>88.1–92.7</td>
<td>99.8</td>
<td>99.8–99.9</td>
<td>87.3</td>
<td>84.8–89.9</td>
</tr>
<tr>
<td>13</td>
<td>746</td>
<td>97.3</td>
<td>96.1–98.6</td>
<td>99.7</td>
<td>99.7–99.8</td>
<td>82.6</td>
<td>79.9–85.3</td>
</tr>
<tr>
<td>14</td>
<td>733</td>
<td>93.0</td>
<td>91.1–95.0</td>
<td>99.7</td>
<td>99.7–99.8</td>
<td>80.4</td>
<td>77.5–83.2</td>
</tr>
</tbody>
</table>

SEN: Sensitivity. SPE: Specificity. PPV: Positive predictive value. 95% CI: Confidence interval.
when new insights regarding breast cancer are obtained, such as developments in diagnosis and treatment procedures.

Despite these limitations, the developed definitions are considered to be efficient in the identification of breast cancer cases, especially in cases for which it is impossible to link a claims database such as the NDB with other sources to obtain additional information. In addition, the patterns selected for these definitions are useful for the identification of breast cancer cases.

In conclusion, the definitions we developed were capable of accurately identifying breast cancer cases from the claims database. We are planning to conduct further work to confirm our results, via multi-institutional studies in Japan.

Acknowledgments The authors wish to thank Yutaka Matsuyama, Ayano Takeuchi, Satoshi Imuro, Yukari Uemura, Kiyoshi Kubota, Nobuhiro Ooba for useful comments regarding this study. The authors also wish to thank NPO Japan Clinical Research Support Unit (J-CRSU) and the St. Luke’s International Hospital for their support.

Conflict of Interest The authors declare no conflict of interest.

Supplementary Materials The online version of this article contains supplementary materials.

REFERENCES