An Algorithm for the Use of Medicare Claims Data to Identify Women with Incident Breast Cancer

Ann B. Nattinger, Purushottam W. Laud, Ruta Bajorunaite, Rodney A. Sparapani, and Jean L. Freeman

Objective. To develop and validate a clinically informed algorithm that uses solely Medicare claims to identify, with a high positive predictive value, incident breast cancer cases.

Data Source. Population-based Surveillance, Epidemiology, and End Results (SEER) Tumor Registry data linked to Medicare claims, and Medicare claims from a 5 percent random sample of beneficiaries in SEER areas.

Study Design. An algorithm was developed using claims from 1995 breast cancer patients from the SEER-Medicare database, as well as 1995 claims from Medicare control subjects. The algorithm was validated on claims from breast cancer subjects and controls from 1994. The algorithm development process used both clinical insight and logistic regression methods.

Data Extraction. Training set: Claims from 7,700 SEER-Medicare breast cancer subjects diagnosed in 1995, and 124,884 controls. Validation set: Claims from 7,607 SEER-Medicare breast cancer subjects diagnosed in 1994, and 120,317 controls.

Principal Findings. A four-step prediction algorithm was developed and validated. It has a positive predictive value of 89 to 93 percent, and a sensitivity of 80 percent for identifying incident breast cancer. The sensitivity is 82–87 percent for stage I or II, and lower for other stages. The sensitivity is 82–83 percent for women who underwent either breast-conserving surgery or mastectomy, and is similar across geographic sites. A cohort identified with this algorithm will have 89–93 percent incident breast cancer cases, 1.5–6 percent cancer-free cases, and 4–5 percent prevalent breast cancer cases.

Conclusions. This algorithm has better performance characteristics than previously proposed algorithms. The ability to examine national patterns of breast cancer care using Medicare claims data would open new avenues for the assessment of quality of care.

Key Words. Breast neoplasm, incidence, sensitivity and specificity, registries, Medicare

The quality of cancer care in the United States is known to be variable, and factors determining quality of cancer care have been insufficiently studied (Hewitt and Simone 1999). The development of methods for using existing

databases to study the quality of cancer care would be a major advance (Hewitt and Simone 2000). Methods to permit the use of Medicare administrative databases to study cancer quality of care would be particularly helpful because about 60 percent of persons diagnosed with cancer are aged 65 and older (Hewitt and Simone 2000), and the Medicare claims data represent a nearly population-based source of data.

With respect to breast cancer specifically, several challenges have been identified in the use of Medicare claims in studying the care provided. The use of inpatient Medicare claims to identify incident breast cancer cases offers excellent specificity but poor sensitivity because 30–40 percent of initial breast cancer operations are done on an outpatient basis (Warren et al. 1999; Warren et al. 1996). Inpatient records are also more likely to identify patients undergoing mastectomy for initial therapy than those undergoing breast-conserving surgery (Warren et al. 1996; Cooper et al. 2000). Compared to inpatient data alone, the use of combined inpatient, outpatient, and physician claims increases sensitivity to 80–90 percent (Freeman et al. 2000; Cooper et al. 1999), but decreases specificity (Warren et al. 1999; Freeman et al. 2000). Because only a small percentage of the female Medicare population develops breast cancer in a given year, even small decreases in specificity lead to large decreases in the positive predictive value (PPV) (Freeman et al. 2000).

Our major goal in the development of this algorithm was to identify a cohort of incident breast cancer patients, whose surgical, medical, and follow-up care could be studied over time. Inherent in this goal was a requirement for a high positive predictive value (PPV), ensuring that a high percentage of the cohort was made up of true breast cancer patients. The requirement for a high

Grant support from the Department of the Army (DAMD17-96-6262).

This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the Applied Research Program, NCI; the Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database.

Address correspondence to Ann B. Nattinger, M.D., M.P.H., Division of General Internal Medicine, Medical College of Wisconsin, FEC Medical Office Building, Suite 4200, 9200 W. Wisconsin Avenue, Milwaukee, WI 53226. Purushottam W. Laud, Ph.D., is with the Department of Biostatistics, Medical College of Wisconsin, Milwaukee; Ruta Bajorunaite, Ph.D., is with the Department of Mathematics, Statistics, and Computer Science, Marquette University, Milwaukee; Rodney A. Sparapani, M. S., is with the Center for Patient Care and Outcomes Research, Medical College of Wisconsin; and Jean L. Freeman, Ph.D., is with the Departments of Medicine and Preventive Medicine and Community Health, University of Texas Medical Branch, Galveston.

PPV was considered more important than the algorithm's sensitivity, particularly for the small percentage (6–7 percent) of women not undergoing initial surgical therapy. However, we also considered important the consistency of the algorithm's sensitivity across subgroups defined by geographic location, age, and type of initial surgery undergone (breast-conserving surgery [BCS] or mastectomy.)

The prior work of the other investigators cited had adequately demonstrated that a relatively simple algorithm (generally consisting of the identification of a claim with a coincident breast cancer diagnosis and operative procedure) would not permit us to achieve our goal. Our strategy was to use an interaction of clinical rationale and statistical analysis in developing the four-step algorithm presented herein.

METHODS

Sources of Data

The key data source for this study was the linked SEER-Medicare database (SEER-Medicare Linked Database 2003). This database links information from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) tumor registries and the Centers for Medicare and Medicaid Services (CMS) Medicare claims data. The population-based SEER registries cumulatively represent about 14 percent of the U.S. population, and include information on incident cancer patients, such as demographics, month and year of diagnosis, extent of disease, and initial treatment undergone. The Medicare files required for this study include the Medicare Provider Analysis and Review (MEDPAR) file, which contains inpatient hospital claims; Outpatient file, which contains claims from institutional outpatient providers including hospital ambulatory surgery centers; the Carrier Claims (previously known as Part B Physician/Supplier File), which contains inpatient and outpatient claims from noninstitutional providers such as physicians, as well as stand-alone ambulatory surgical centers; and the Denominator file, which contains beneficiary demographic information and Medicare entitlement and enrollment information. About 94 percent of the SEER registry patients aged 65 and older were successfully linked with their Medicare claims (Potosky et al. 1993). An additional data source was a 5 percent random sample of Medicare beneficiaries residing in the SEER geographic areas, including an indicator for whether the individual linked to the SEER database. When SEER subjects are removed from this sample, it represents nearly a population-based random sample of cancer-free

control subjects residing in SEER areas. This study was approved by the Medical College of Wisconsin Human Subjects Research Review Committee.

Training and Validation Datasets

Training Set: Incident Breast Cancer Cases. A cohort of women aged 65 or older at the time of diagnosis of breast cancer in 1995 (according to SEER) was developed. Cases were excluded if the diagnosis was made only at autopsy or by death certificate. Subjects were required to meet the following criteria for the period from January 1995 to March 1996: eligibility for Medicare Parts A and B, not in a Medicare HMO, and to be alive. Eligibility through the first quarter of 1996 was required to capture Medicare treatment information for patients who were diagnosed near the end of 1995, but treated early in 1996. These criteria resulted in a cohort of 7,700 women, whose 1995 Medicare claims comprised the training set for incident breast cancer cases.

Training Set: Cancer-Free Subjects. From the 5 percent random sample of Medicare beneficiaries who resided in SEER areas but who did not link to SEER registry for an incident cancer between 1973 and 1995, a cancer-free cohort was developed. These 71,752 women were required to meet the same eligibility criteria as the breast cancer cases with respect to Medicare eligibility and survival. The 1995 claims of these women comprised the "cancer-free" training set.

Training Set: Other Cancer Cases. Again using the 5 percent random sample and the same eligibility criteria, a cohort was constructed of 4,501 women who were diagnosed with a cancer other than breast cancer between 1973 and 1995. The 1995 claims of these women comprised the "other cancer" training set.

Prevalent Breast Cancer Cases. For the purposes of this article, we use the term prevalence to indicate cancer cases diagnosed prior to the index year and not including the incident cases diagnosed in the index year. From the SEER-Medicare linked data, a cohort was developed of 48,631 women who had breast cancer between 1973 and 1994, according to SEER, and who were alive and eligible for Medicare Parts A and B and not in an HMO during 1995. The 1995 claims for these women were not used to train the algorithm per se, but were used to assess the impact of prevalent cases on the algorithm's specificity.

Validation Sets. Using the same selection criteria as described above for the year 1995, four analogous sets of claims were constructed for calendar year 1994. When evaluating a predictive algorithm, it is important that the validation set be independent of the training set. We defined the training sets to be comprised of claims from 1995, while the validation sets were comprised of

independent claims from 1994. We recognized the possibility that some of the individuals generating the claims for the 1995 training set might also have generated claims for the 1994 validation set, particularly among the cancer-free and other cancer groups. In assessing the frequency of such overlap, though, we determined that only 5.5 percent of the individuals whose claims were part of the algorithm's training for steps 2 or 3 (described below under "Algorithm Development") also contributed claims to the validation set at those steps.

Algorithm Development

The algorithm was developed using the 1995 training sets. In constructing the algorithm, consideration was given not only to the presence or absence of breast cancer diagnosis or procedure codes, but also to other related codes (such as historical codes and radiation codes) that might improve the prediction of a case. In addition, variables were evaluated indicating whether a code was in a primary or secondary position on a given claim, and how frequently the code occurred. The algorithm development effort involved an iteratively processed interaction between clinical insight and statistical analysis. The codes actually used in the algorithm are summarized in Table 1.

A four-part algorithm was developed (Figure 1). The input to the algorithm consists of Medicare claims of all women aged 65 and older who were alive and eligible for Medicare Part A and Part B in some index year, including claims for the following three months.

Step 1. Referred to as the "screen," requires that a potential case have both a breast cancer diagnosis code and a breast cancer procedure code (not necessarily on the same claim) anywhere in the Medpar, Outpatient, or Carrier Claims records (see Table 1). Only subjects satisfying this screening step are retained for further consideration.

Step 2. Directly includes subjects with a high likelihood of being a case. To be classified as a case based on this step, the subject must meet *both* of the following criteria:

1. A mastectomy claim in any file *or* a lumpectomy or partial mastectomy claim in any file followed by at least one Outpatient or Carrier claim for radiotherapy with a breast cancer diagnosis.
2. At least two Outpatient or Carrier claims on different dates containing breast cancer as the primary diagnosis.

Subjects who pass step 2 are classified as possible incident cases, and proceed to step 4. Subjects who are not classified as cases at step 2 go to step 3.

Table 1: Claims Codes Used as Possible Predictors of Breast Cancer

<i>Diagnosis or Procedure</i>	<i>ICD-9-CM Codes</i>	<i>CPT/HCPCS Codes (procedures only)</i>
Breast cancer*	174–174.9	
Carcinoma-in-situ (breast)*	233.0	
History of breast cancer	V10.3	
Tumor in breast of uncertain nature	238.3, 239.3	
Secondary cancer to breast	198.2, 198.81	
Other cancer	140–173.9 175–195.8 197–199.1 (excluding 198.2, 198.81) 200–208.91 230–234.9 (excluding 233.0, 232.5) 235–239.9 (excluding 238.3, 239.3)	
Biopsy [†]	85.1–85.19	19000, 19001, 19100, 19101, 19110, 19112
Lumpectomy ^{†,§}	85.20–85.21	19120, 19125, 19126
Partial mastectomy ^{†,§}	85.22–85.23	19160, 19162
Lymph node dissection [†]	40.3	38740, 38745, 38525
Mastectomy ^{†,§}	85.33–85.48	19180–19255
Radiation therapy	92.2–92.29	77400–77499 77520–77525 77750–77799

*Diagnoses for Step 1;

[†]Procedures for Step 1;[§]Procedures for “Surgery” variable in Step 3.

Step 3. This step of the algorithm applies to all potential cases that passed the screen (step 1), but were not directly included at step 2. In practice, this step differentiates primary breast cancer cases from women undergoing lumpectomy or partial mastectomy for benign disease or for another cancer that had metastasized to the breast. Some such patients without primary breast cancer have claims with erroneous primary breast cancer diagnosis codes, and therefore pass the step 1 screen. To develop step 3 of the algorithm, logistic regression methods were employed. A model was developed predicting an incident breast cancer case from about 50 indicator variables representing the presence of various billing codes (diagnostic or procedural) or combinations of codes that were thought clinically to have possible usefulness. Complete details regarding this model are omitted in the interest of space, and because

Figure 1: Schematic Representation of a Four-Step Algorithm for Identifying Incident Breast Cancer Cases from Medicare Claims

The initial cohort on which the algorithm operates is comprised of women aged 65 and older on January 1 of the index year, who are alive and eligible for Medicare parts A and B, and not in an HMO, from January 1 of the index year through March 31 of the following year. Further details of each step are provided in the text.

such details do not assist in assessing the performance of the final algorithm. The final parsimonious logistic model included only the following four dichotomous factors:

Surgery. This variable is positive (i.e., set to a value of 1 in the regression equation) if one or more lumpectomy, partial mastectomy, or mastectomy codes are found in any file (see Table 1). Otherwise, the variable is negative (set to a value of zero).

Single Claim. This variable is positive (i.e., set to a value of 1) if a woman with lumpectomy or partial mastectomy claim in any file had only one month in which a claim contained a primary breast cancer or a breast carcinoma-in-situ diagnosis. Otherwise, this variable is negative (i.e., set to 0).

Other Cancer. This variable is positive (i.e., set to 1) if an "other cancer" code is found as a primary diagnosis in one or more claims from any file. Otherwise, this variable is set to 0.

Secondary Cancer to Breast. This variable is positive (i.e., set to 1) if a code for secondary cancer to breast is found in one or more Outpatient or Carrier claims. Otherwise, this variable is set to 0.

Because all the factors in the model are binary variables, it is not necessary for a user of the algorithm to use the regression equation to classify a case as positive or negative. Once the values of the four variables have been determined, subjects can be ruled in if they have one of three combinations of the variables. These combinations are (1) the "surgery" variable is positive and the other three variables are negative, (2) the "surgery" variable is positive, the "other cancer" variable is positive, and the other two variables are negative, or (3) the "surgery" variable is positive, the "secondary cancer to breast" variable is positive, and the other two variables are negative. With all other combinations, the subject is declared not to be a breast cancer case.

Step 4. This step of the algorithm is the step to remove prevalent breast cancer cases. This step uses three prior years of claims of subjects classified as a case in step 2 or step 3. Such subjects are removed if they have a claim in prior years 1992–1994 (1991–1993 for the validation cohort) that was either positive for step 1 (the screening step) of the algorithm, or contained a diagnosis of prior history of breast cancer. Women younger than age 68 at diagnosis did not have three full years of claims for review, but as many years as were available were used. This strategy results in the removal of most prior cases, but also a number of incident cases (Table 2, rows 6 and 7).

Determination of Incident Breast Cancer Cases

The initial approach was to consider the "gold standard" for defining an incident breast cancer case to be SEER. However, while conducting this study, it was determined that subjects meeting the criteria from step 2 above had a very high likelihood of being a case according to SEER. A small number of cancer-free control subjects also met the step 2 high likelihood criteria. After manual inspection of the claims for these subjects, the likelihood of these being

Table 2: Performance of Algorithm for Years 1994, 1995

	1994				1995			
	Breast Cancer	Other Cancer	Cancer Free	Prior Breast Cancer	Breast Cancer	Other Cancer	Cancer Free	Prior Breast Cancer
Cohort size	7,607	4,360	72,106	43,851	7,700	4,501	71,752	48,631
Step 1: Screen positive	7,244	15	74	1,777	7,363	21	71	1,953
Step 2: High likelihood	5,693	1	19	564	5,837	1	17	624
Step 2: Not high likelihood	1,551	14	55	1,213	1,526	20	54	1,329
Step 3: Number positive	796	3	7	615	830	4	9	618
Incident or prevalent cases	6,489	4	26	1,179	6,667	5	26	1,242
(Positive after Step 2 or 3)								
Step 4: Incident breast cancer cases	6,094	3	20	287	6,180	3	24	318
Number correctly (incorrectly)	6,094	4,357	72,086	43,564	6,180	4,498	71,728	48,313
classified, GS = SEER	(1,513)	(3)	(20)	(287)	(1,520)	(3)	(24)	(318)
% Correctly classified, GS = SEER	80.11	99.93	99.97	99.35	80.26	99.93	99.97	99.35
Number correctly (incorrectly)	6,094	4,358	72,101	43,564	6,180	4,499	71,744	48,313
classified, GS = SEER+HL	(1,513)	(2)	(5)	(287)	(1,520)	(2)	(8)	(318)
% Correctly classified, GS = SEER+HL	80.11	99.95	99.99	99.35	80.26	99.96	99.99	99.35

GS = Gold Standard.

SEER = Surveillance, Epidemiology, and End Results gold standard (defined in text).

SEER+HL = SEER plus High Likelihood gold standard (defined in text).

incident breast cancer cases seemed extremely high (based on the pattern and number of claims with breast cancer diagnosis over time, radiotherapy claims, etc.). Therefore, in the results section two gold standards for defining an incident case of breast cancer are reported. The first is termed the "SEER" gold standard, which is defined solely by whether a subject linked to the SEER registry in that year as a breast cancer case. The second gold standard is termed the "SEER plus High Likelihood" gold standard and consists of cases identified by a SEER registry as well as control subjects identified by the two criteria of step 2 above. Our belief is that these cases might be in the group of about 6 percent of SEER subjects who did not successfully link with Medicare files (Potosky et al. 1993) or they might be due to a patient moving into a SEER area shortly after breast cancer diagnosis.

Computation of PPV

The estimates of sensitivity and specificity were converted to an estimate of the positive predictive value (PPV) using Bayes Theorem, as

$$PPV = \frac{\pi_B \Pr(+|B)}{\pi_B \Pr(+|B) + \pi_O \Pr(+|O) + \pi_{PB} \Pr(+|PB) + \pi_N \Pr(+|N)}$$

where π_B , π_O , π_{PB} , π_N represent the incidence of breast cancer, incidence and prevalence of other cancers, prior breast cancer, and no cancer, respectively, in the study population. Based on SEER data, these were estimated to be 0.005, 0.07, 0.03, and 0.895, respectively. Confidence intervals for the PPVs were estimated using Fieller's method (Fieller 1940; Steffens 1971).

RESULTS

When applied to the training cohort, this algorithm had excellent specificity, and moderate sensitivity (Tables 2 and 3). Of the initial breast cancer cohort, about 5 percent of the subjects were not detected by step 1, about 9 percent were not retained by step 3, and a further 6 percent had a prior year code for breast cancer diagnosis or history thereof at step 4. This left an overall sensitivity of 80 percent. The specificity was excellent, at well over 99.9 percent.

Algorithm Validation

The validation of this algorithm was carried out in the 1994 cohorts (Tables 2 and 3). The algorithm's performance was similar to the training year. The specificity of the algorithm remained well above 99.9 percent. Using the

Table 3: Positive Predictive Values (%) of an Algorithm for Using Medicare Claims to Identify Incident Breast Cancer

	1994 (Validation Year)		1995 (Training Year)	
	Gold Standard SEER+HL	Gold Standard SEER	Gold Standard SEER+HL	Gold Standard SEER
Sensitivity	80.11	80.11	80.26	80.26
Specificity				
Other cancer	99.95	99.93	99.96	99.93
Cancer-free	99.99	99.97	99.99	99.97
Prior breast cancer	99.35	99.35	99.35	99.35
Overall	99.97	99.95	99.97	99.95
Positive Predictive Value (PPV)	93.24	89.05	92.46	88.10
95% Confidence Interval for PPV	91.66–94.87	86.66–91.57	90.70–94.30	85.60–90.74
Algorithm Cohort Composition				
Incident breast cancer	93.24	89.05	92.46	88.10
Other cancer	0.75	1.07	0.72	1.02
Cancer-free	1.44	5.52	2.30	6.57
Prior breast cancer	4.57	4.36	4.52	4.31

SEER = Surveillance, Epidemiology, and End Results gold standard (defined in text).

SEER+HL = SEER plus High Likelihood gold standard (defined in text).

stricter “SEER” gold standard, the PPV was 89 percent. Using the “SEER plus high likelihood” gold standard, the PPV was 93 percent.

Using the PPVs for the four cohorts, the expected composition of a cohort developed using this algorithm can be determined (Table 3). In the validation year, the vast majority of cases selected by the algorithm are incident breast cancer cases. About 1 percent are other cancer cases. About 4–5 percent of cases selected by the algorithm are prevalent breast cancer cases. A substantial minority of the prevalent breast cancer cases was diagnosed according to SEER in the three months prior to the start of the 1994 year. If one were willing to tolerate a three-month error in date of diagnosis (i.e., cases diagnosed according to SEER in the last three months of 1993 are not counted against the algorithm’s specificity for 1994), the percent of prevalent cases in the 1994 algorithm cohort would decrease from 4.6 percent to about 3 percent, and the percent considered incident cases would increase accordingly. With respect to the composition of the algorithm cohort, the percentage of cancer-free patients in the validation year varies from 1.4 percent to 5.5 percent depending on which gold standard is used.

We performed a sensitivity analysis of the specificity gain associated with examining prior claims in step 4 for differing numbers of years. Of the 48,631

prior breast cancer cases with 1995 claims, only 1,242 were positive after step 2 or 3 of the algorithm. Examining prior claims for one year back in step 4 would have removed 58.6 percent of those cases. Going back two, three, or four years, respectively, removed 69.3 percent, 74.4 percent, and 76.5 percent of the 1,242 cases. The specificity gain from applying step 4, however, is associated with a sensitivity loss (loss of index year true incident cases who met the criteria for removal at step 4). The percentage of true incident cases retained when applying step 4 going back one, two, three, or four years was 95.4 percent, 93.9 percent, 92.7 percent, and 92.3 percent respectively.

Algorithm Sensitivity by Patient Characteristics

The algorithm sensitivity by selected patient characteristics is presented in Table 4. The algorithm sensitivity is lower for women with stage 4 and unknown stage disease at presentation, but there are relatively few such patients in any given year. Women are well represented up to age 84, but there is a decline in sensitivity for the women aged 85 and older. The sensitivity is consistent across the different SEER geographic regions. With respect to initial treatment, the algorithm fails to identify women who did not undergo initial surgery according to SEER, but identifies equally well women who underwent mastectomy and those who underwent BCS. Women who underwent lymph node dissection or radiotherapy according to SEER are somewhat overrepresented compared to those who did not.

DISCUSSION

We propose a 4-step algorithm for the use of Medicare claims data to identify women with surgically treated incident breast cancer. This algorithm has a sensitivity of about 80 percent overall, with a sensitivity of 82–87 percent for stages 1 and 2 disease. The algorithm has a specificity above 99.9 percent, and a positive predictive value of 89 percent, using a SEER gold standard. The PPV is greater than 93 percent based on the SEER plus High Likelihood gold standard.

The algorithm development process described herein illustrates several major issues with respect to the use of Medicare claims to identify breast cancer cases. One is the relationship of specificity to positive predictive value. Because only a minority of women, even in the Medicare age group, develop breast cancer in a given year, an exceedingly high specificity (>99.9 percent) is necessary to have a positive predictive value of 90 percent. The dramatic decline in PPV that occurs with only small decreases in specificity can be seen

Table 4: Sensitivity of the Algorithm for 1994, by Selected Patient Characteristics

<i>Subgroup Categories*</i>	<i>Number (%) Identified by Algorithm</i>	<i>Number in SEER Cohort</i>	<i>Odds Ratio** with 95% Confidence Interval</i>
Overall	6,094 (80.1)	7,607	
Modified AJCC Stage			
In situ	665 (72.8%)	914	0.90 (0.81, 1.00)
I	2,818 (81.9%)	3,441	1.04 (0.97, 1.11)
IIa	1,299 (85.3%)	1,522	1.08 (1.00, 1.18)
IIb	506 (89.6%)	565	1.13 (1.00, 1.28)
II, NOS	52 (89.7%)	58	1.12 (0.77, 1.63)
IIIa	169 (88.0%)	192	1.10 (0.89, 1.36)
IIIb	150 (73.9%)	203	0.92 (0.74, 1.14)
IV	89 (51.4%)	173	0.64 (0.49, 0.82)
Unknown	346 (64.2%)	539	0.79 (0.69, 0.91)
Age			
65–74	3,379 (83.5%)	4,047	1.09 (1.02, 1.17)
75–84	2,222 (79.8%)	2,784	0.99 (0.93, 1.07)
85+	493 (63.5%)	776	0.77 (0.69, 0.87)
SEER Area			
Atlanta	350 (78.7%)	445	0.98 (0.85, 1.13)
Connecticut	945 (82.5%)	1,146	1.03 (0.94, 1.14)
Detroit	1,044 (86.7%)	1,211	1.09 (1.00, 1.20)
Hawaii	132 (84.6%)	156	1.06 (0.84, 1.34)
Iowa	817 (78.0%)	1,047	0.97 (0.88, 1.07)
Los Angeles County	880 (81.6%)	1,078	1.02 (0.93, 1.13)
New Mexico	214 (80.1%)	267	1.00 (0.83, 1.20)
San Francisco/Oakland	515 (76.3%)	675	0.95 (0.84, 1.07)
San Jose-Monterey	266 (78.0%)	341	0.97 (0.83, 1.15)
Seattle-Puget Sound	675 (74.4%)	907	0.92 (0.83, 1.02)
Utah	256 (76.6%)	334	0.95 (0.81, 1.13)
Surgery			
None	83 (27.7%)	300	0.34 (0.26, 0.43)
Lumpectomy/partial mastectomy	2,746 (83.0%)	3,310	1.06 (0.99, 1.14)
Mastectomy	3,264 (81.8%)	3,990	1.05 (0.98, 1.12)
Unknown	1 (14.3%)	7	0.18 (0.02, 1.45)
Lymph Node Dissection			
Yes	4,393 (85.3%)	5,153	1.23 (1.14, 1.32)
No	1,701 (69.5%)	2,447	0.82 (0.76, 0.88)
Unknown	0 (0.0%)	7	0.00
Radiotherapy			
Yes	2,101 (90.2%)	2,329	1.19 (1.11, 1.28)
No	3,930 (75.6%)	5,198	0.84 (0.78, 0.90)
Unknown	63 (78.8%)	80	0.98 (0.71, 1.37)

AJCC = American Joint Commission on Cancer.

SEER = Surveillance, Epidemiology, and End Results.

*All categories are based on the SEER data.

**This odds ratio represents the ratio of odds in the algorithm cohort to odds in the SEER cohort for that category. For example, the value of 1.04 for stage I means that such cases are slightly overrepresented in the algorithm cohort.

Table 5: Comparison of Algorithms Using Medicare Claims to Identify Breast Cancer Subjects in Tumor Registries

<i>Algorithm</i>	<i>Sensitivity (%)</i>	<i>Specificity (%)</i>	<i>PPV (%)</i>	<i>Comments</i>
McClish et al. 1997	83.0			Only sensitivity was assessed.
Cooper et al. 1999	82.0			Only sensitivity was assessed.
Warren et al. 1999	76.2	99.3	36.3	Inpatient+physician claims.
Warren et al. 1999	57.0	99.9	91.3	Inpatient claims only.
Freeman et al. 2000	90.0	99.86	70.0	Inpatient, outpatient, and physician claims. PPV 67% if prevalent claims included.
Current study	80.1	99.95	89.0	Inpatient, outpatient, and physician claims. Gold standard is SEER.
Current study	80.1	99.97	93.2	Inpatient, outpatient, and physician claims. Gold standard is SEER + High Likelihood.

PPV = Positive Predictive Value.

SEER = Surveillance, Epidemiology, and End Results.

by comparing the results of this algorithm with prior proposed algorithms (Table 5). Given that the procedures used to treat breast cancer may also be used to identify or treat benign breast disease, and given occasional inaccuracies in the use of a breast cancer diagnosis, it is challenging to reach the necessary level of specificity.

A major goal of this algorithm was to maintain a high specificity while including cases treated in the ambulatory surgical setting. This algorithm achieves a PPV similar to that reported by Warren and colleagues (Warren et al. 1999) for inpatient claims, while providing improved sensitivity (Table 5). Although the sensitivity is not as high as that reported by Freeman and colleagues (Freeman et al. 2000), the PPV is much higher.

Another major issue with this and prior algorithms is the presence of prevalent cases. Because women with breast cancer often live for many years, the number of prevalent cases in a dataset greatly exceeds the number of incident cases. Women with prevalent disease undergo at times the same breast procedures as women with initial disease to diagnose or rule out recurrent or new breast disease, and also may carry diagnostic codes of primary breast cancer for years after initial disease. Since local disease recurrences occur most frequently within the first few years after diagnosis, our approach was to assume that algorithm-identified cases with a history of breast cancer within the prior three years had recurrent disease. This led to a decrease in sensitivity from about 85 percent to 80 percent, but maintained the high specificity of the algorithm.

In attempting to maximize the PPV of the algorithm, we accepted a moderate sensitivity of about 80 percent. Therefore, this algorithm may have limited utility for determining breast cancer incidence. The key uses for this algorithm are likely to be for aspects of care not well captured by SEER or other state tumor registries. The study of survivor care, for example, studies of mammography (Schapira, McAuliffe, and Nattinger 2000) or other health care utilization and physician care (Nattinger et al. 2002) among survivors, is well suited to claims analysis. Patterns of care studies with respect to geographic variation and rural areas not well represented by SEER appear feasible given the consistency of the algorithm in different geographic areas. Studies might examine pre-morbid care for older breast cancer patients, such as use of mammography or other preventive care interventions. Although some of the studies mentioned could be performed using the limited number of available linked tumor registry–Medicare databases, the need for greater geographic representation or larger sample sizes might favor the use of Medicare-derived samples. Given that almost half of all breast cancer cases occur in women aged 65 and older, the algorithm could be applied to 100 percent state Medicare databases for identifying providers with possible quality problems, such as low levels of medical oncology consultation, poor follow-up care, and poor preventive care practices. An algorithm that is less than perfect may still provide a valid assessment of patterns of care (Kahn et al. 1996).

A limitation we encountered is the fact that about 5 percent of women identified by SEER as having an incident breast cancer, and who linked to the Medicare claims data, did not even pass the screening step. Based on Table 4 it appears that some of these women do not undergo initial surgical therapy. Perhaps some women undergo surgery but are covered by employer-based insurance, which pays for their care in preference to Medicare. In any event, this problem does cause a limitation on the sensitivity that can be achieved by the algorithm, even if steps 2, 3, and 4 could be further optimized. Another limitation is that women who underwent radiotherapy are somewhat over-represented compared to those who did not, limiting the ability to use this algorithm to study patterns of care for radiotherapy.

We are not able to state which of the two “gold standards” represents a more accurate definition of an incident breast cancer case. Although the SEER tumor registry program has excellent case ascertainment, all registry programs likely miss occasional cancer cases. In the case of this study, an incident cancer patient could also have been classified as a cancer-free control subject due to failure to link with the Medicare beneficiary files, or due to moving into a SEER area shortly after disease diagnosis. For these reasons, we developed

and presented the "SEER plus High Likelihood" gold standard, which followed a decision rule created initially by manual inspection of the claims histories for certain control subjects who seemed to have a high likelihood of having breast cancer. We were convinced that these subjects likely had incident breast cancer by the lack of prior claims suggesting prevalent disease, and by the multiple claims during the training year that consistently suggested an operation for breast cancer (surgical claims, pathology claims, anesthesiology claims, etc.). Since we did not have access to patient identifiers or charts, we could not confirm that these patients had breast cancer. However, Warren and colleagues (1999) have previously demonstrated that some cases identified by their algorithm using Medicare claims actually had breast cancer but failed to link to SEER when the linkage was conducted. In addition, the number of high-likelihood cases identified by our algorithm within the 5 percent control sample is very close to the number that one would expect given a 94 percent linkage rate between SEER and Medicare. For example, in 1994, a 6 percent failure to link would translate into 456 unlinked breast cancer cases. We would expect 5 percent of these (23 cases) to be found in the 5 percent control sample. We would further expect the high-likelihood definition to identify 75 percent (17 cases). In fact, the high-likelihood definition did identify 19 cases in the 5 percent control cohort that year (Table 2), very close to the expected number.

As has been shown in a number of other disease areas, Medicare claims data offer unique advantages for cancer quality of care and health services research (Hewitt and Simone 1999, 2000; McNeil 2001). These data are essentially population-based, and minimize selection bias with respect to geographic region, urban versus rural location, and socioeconomic status. Each of these factors is an important predictor of cancer treatment, a fact that limits analyses of databases from more restricted populations (Nattinger et al. 1992; Guadagnoli et al. 1998; Gilligan et al. 2002). The possibility of using Medicare data more widely to assess patterns of cancer practice and related outcomes offers a potential that is worthy of further exploration.

REFERENCES

- Cooper, G. S., Z. Yuan, K. C. Stange, L. K. Dennis, S. B. Amini, and A. A. Rimm. 2000. "Agreement of Medicare and Tumor Registry Data for Assessment of Cancer-Related Treatment." *Medical Care* 38 (4): 411-21.
- . 1999. "The Sensitivity of Medicare Claims Data for Case Ascertainment of Six Common Cancers." *Medical Care* 37 (5): 436-44.
- Fieller, E. C. 1940. "The Biological Standardization of Insulin." *Journal of the Royal Statistical Society* 7 (supplement): 1-64.

- Freeman, J., D. Zhang, D. Freeman, and J. Goodwin. 2000. "An Approach to Identifying Incident Breast Cancer Cases Using Medicare Claims Data." *Journal of Clinical Epidemiology* 53 (6): 605-14.
- Gilligan, M. A., R. T. Kneusel, R. G. Hoffmann, A. L. Greer, and A. B. Nattinger. 2002. "Persistent Differences in Sociodemographic Determinants of Breast Conserving Treatment Despite Overall Increased Adoption." *Medical Care* 40 (3): 181-9.
- Guadagnoli, E., C. L. Shapiro, J. C. Weeks, J. H. Gurwitz, C. Borbas, and S. B. Soumerai. 1998. "The Quality of Care for Treatment of Early Stage Breast Carcinoma. Is It Consistent with National Guidelines?" *Cancer* 83 (2): 302-9.
- Hewitt, M., and J. V. Simone, eds. 1999. *Ensuring Quality of Cancer Care*. Washington, DC: National Academy Press.
- . 2000. *Enhancing Data Systems to Improve Quality of Cancer Care*. Washington, DC: National Academy Press.
- Kahn, L. H., J. Blustein, R. R. Arons, R. Yee, and S. Shea. 1996. "The Validity of Hospital Administrative Data in Monitoring Variations in Breast Cancer Surgery." *American Journal of Public Health* 86 (2): 243-5.
- McClish, D. K., L. Penberthy, M. Whittemore, C. Newschaffer, D. Woolard, C. E. Desch, and S. Retchin. 1997. "Ability of Medicare Claims Data and Cancer Registries to Identify Cancer Cases and Treatment." *American Journal of Epidemiology* 145 (3): 227-33.
- McNeil, B. J. 2001. "Shattuck Lecture: Hidden Barriers to Improvement in the Quality of Care." *New England Journal of Medicine* 345 (22): 1612-20.
- Nattinger, A. B., M. S. Gottlieb, J. Veum, D. Yahnke, and J. S. Goodwin. 1992. "Geographic Variation in the Use of Breast-Conserving Treatment for Breast Cancer." *New England Journal of Medicine* 326 (17): 1102-7.
- Nattinger, A. B., M. M. Schapira, J. L. Warren, and C. C. Earle. 2002. "Methodologic Issues in the Use of Administrative Claims Data to Study Surveillance after Cancer Treatment." *Medical Care* 40 (8): IV-69-74.
- Potosky, A. L., G. F. Riley, J. D. Lubitz, R. M. Mentnech, and L. G. Kessler. 1993. "Potential for Cancer Related Health Services Research Using a Linked Medicare-Tumor Registry Database." *Medical Care* 31 (8): 732-48.
- Schapira, M. M., T. L. McAuliffe, and A. B. Nattinger. 2000. "Underutilization of Mammography in Older Breast Cancer Survivors." *Medical Care* 38 (3): 281-9.
- SEER-Medicare Linked Database. 2003. "National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) tumor registries and the Centers for Medicare and Medicaid Services (CMS) Medicare claims data" [accessed July 1, 2003]. Available at <http://healthservices.cancer.gov>.
- Steffens, F. E. (1971). "On Confidence Sets for the Ratio of Two Normal Means." *South African Statistical Journal* 5 (2): 105-13.
- Warren, J. L., E. Feuer, A. L. Potosky, G. F. Riley, and C. F. Lynch. 1999. "Use of Medicare Hospital and Physician Data to Assess Breast Cancer Incidence." *Medical Care* 37 (5): 445-6.
- Warren, J. L., G. F. Riley, A. M. McBean, and R. Hakim. 1996. "The Use of Medicare Data to Identify Incident Breast Cancer Cases." *Health Care Financing Review* 18 (1): 237-46.