AD_____

Award Number: DAMD17-97-1-7095

TITLE: The Elevated Breast Cancer Mortality in the Northeastern U.S. is Secondary to Poorer Survival Rather Than Increased Incidence

PRINCIPAL INVESTIGATOR: James S. Goodwin, M.D.

CONTRACTING ORGANIZATION: University of Texas Medical Branch at Galveston
Galveston, Texas 77555-0136

REPORT DATE: December 1999

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1999	Final (1 Aug 97 - 15 Nov 99)

4. TITLE AND SUBTITLE
The Elevated Breast Cancer Mortality in the Northeastern U.S. is Secondary to Poorer Survival Rather Than Increased Incidence.

5. FUNDING NUMBERS
DAMD17-97-1-7095

6. AUTHOR(S)
James S. Goodwin, M.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Texas Medical Branch at Galveston
Galveston, Texas 77555-0136

E-MAIL:
jsgoodwi@utmb.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Mortality rates for breast cancer are approximately 25% higher in the northeast U.S. than the south or west. Virtually all previous investigations into this phenomenon have assumed that the elevated mortality rates reflect elevated incidence, and this putative elevation in incidence has been attributed to a number of factors. We hypothesized that the geographic variation in breast cancer mortality was secondary to differences in survival with breast cancer as well as differences in incidence. We proposed to test this by identifying women aged $\geq$ 70 years newly diagnosed with breast cancer in 1991 in different regions of the country using Medicare charge data and assessing their survival. Initially we developed algorithms to determine incident breast cancer and date of diagnosis from the Medicare data and validating them using the SEER-Medicare linked data base. We found that an algorithm incorporating data from hospital inpatient, hospital outpatient, and physician services claims produced levels of sensitivity and specificity > 90%. However, the positive prediction value was low (67 to 70 percent), precluding use of Medicare data alone to examine survival from breast cancer. Our current work is using SEER-Medicare linked data to examine geographic variation in incidence, survival and mortality of women with breast cancer.

14. SUBJECT TERMS
Breast cancer survival; geographic variation; incidence; older women; Medicare data

15. NUMBER OF PAGES
45

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

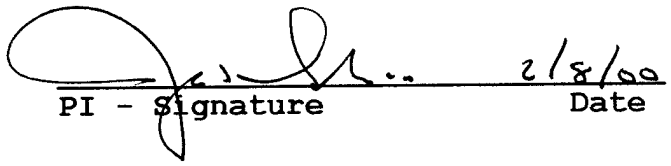
X In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



PI - Signature

2/8/00

Date

Table of Contents

5. INTRODUCTION

The well-described elevation in death rate from breast cancer in the northeastern U.S. has universally been ascribed to an increased breast cancer incidence in that area of the country compared to the rest of the U.S. The fundamental hypothesis driving this proposal is that differences in survival, not incidence, are responsible at least in part for the elevated breast cancer death rate in the northeast.

The existence of the SEER-Medicare linked data base allows us to develop and validate algorithms for incident breast cancer in older women. Such algorithms would allow for the generation of breast cancer incidence and survival estimates for women aged 65 and older at a state and local level; using only Medicare data, which would allow us to directly test this hypothesis of the study.

6. BODY

Task: apply for and obtain Medicare claims for all women with breast cancer diagnosis or procedure in 1991.

Such a file was purchased from HCFA. It contains a finder file of >100,000 women with a breast cancer diagnosis and procedure, plus their Part A and B charge data from 1990 through 1996.

Task: Test the validity and refine initial algorithms against SEER-Medicare linked files.

This work is reported in detail in the appended manuscript by J. Freeman, et al, in press at J. Clin. Epidemiology. We found that none of the algorithms for breast cancer were of sufficient specificity to allow us to accurately determine breast cancer incidence from Medicare data alone. Accordingly, we modified the aims of the remainder of the study in order to examine geographic variation in breast cancer mortality, incidence and survival using the linked Medicare-SEER data.

Task: Calculate breast cancer incidence by state; calculate survival of the cohort by state.

Because we could not use the national Medicare data due to the poor positive predictive value of our algorithm, we modified this task to calculate incidence, survival with and mortality from breast cancer by health service area (HSA) in the SEER areas.

Health Service Areas (HSAs) are aggregations of counties and independent cities based on a cluster analysis of where Medicare patients obtained routine hospital care in 1988 (1). They are a good tool to examine variations in medical practice and outcomes across small areas (1).

We calculated breast cancer mortality, incidence and survival for each of the 73 HSAs in the 9 SEER areas. We present below the mean, median, 25^{th} and 75^{th} percentiles for breast cancer mortality, incidence and survival. The survival was expressed as five year cancer-specific death rate, calculated from the SEER data. Incidence was also from SEER. Mortality was calculated from 1990 Vital Statistics data. All data were for women age 65-74 or 75+ for the years 1985-90 (so that there would be five years of follow-up).

Table 1. Incidence, Mortality and Cancer-specific Death Rate for Breast Cancer in 73 Health Service Areas Contained Within the SEER Sites.

	65-74 Years			75+ Years		
	Incidence	Mortality	Five Year Cancer-Specific Death Rate (%)	Incidence	Mortality	Five Year Cancer-Specific Death Rate (%)
Mean	490	109	13.1%	531	139	14.7%
Median	510	109	12.6%	579	141	14.6%
75% Q3	563	127	9.7%	626	172	12.4%
24% Q1	432	86	15.5%	463	112	18.2%

Incidence and mortality were expressed as per 100,000 adult women. Five year cancer-specific death rate was percent breast cancer deaths at five years.

There is variation among the HSAs in all three measurements: incidence, mortality and five year cancer-specific death rate. For breast cancer incidence, the interquartile range represents approximately 25-30% of the median in the two age categories; for mortality the interquartile range is approximately 40% of the median, while for five year cancer-specific death rate the interquartile range is approximately 50% of the median in the 65-74 year olds and 40% in the 75+ group. Thus, survival with breast cancer varies by geographic area as much as does incidence and mortality.

These variations are stable over time; for all three measurements. HSAs with high or low values for incidence, survival or cancer-specific death rate in 1985-87 also tended to have similar rates in 1988-90. The most compelling evidence that variation in survival contributes to variation in mortality rate is found in correlations between five year survival rate (calculated as 1 – cancer-specific death rate) and mortality rate for breast cancer. In a partial correlation controlling for disease incidence, the coefficient of correlation between mortality rate and survival was $r = 0.37$ ($p = 0.001$) for women 65-74 and $r = 0.52$ ($p = 0.0001$) for women aged 75+.

It is important to remember that mortality rate and survival rate come from two entirely different sources of data. Mortality rate by HSA is derived from U.S. Vital Statistics data, while the survival rate was calculated for incident cases identified by SEER in 1985-90 and followed for at least five years. This provides strong support for the underlying hypothesis of this proposal, that geographic variations in survival from breast cancer contribute to the geographic variation in breast cancer mortality.

In our calculations of survival we used breast cancer-specific survival (or 100% minus breast cancer-specific five year death rate). We used that figure rather than total survival because breast cancer mortality death rates in the Vital Statistics data measure only cases where breast cancer is listed as the underlying cause of death.

Geographic Variation in Treatment for Breast Cancer

We present below evidence that the receipt of definitive treatment varies by geographic area, using SEER data. Table 2 presents the percentage of women aged 65-74 with local or regional stage breast cancer who receive BCS without radiotherapy. It also presents the percent of women aged 65-74 years who received

neither axillary node dissection nor radiotherapy for local or regional stage breast cancer diagnosed from 1983 to 1995. In both cases there is a greater than two-fold variation among the SEER areas.

The Initial Treatment of Older Women with Breast Cancer is not Improving Over Time

Table 3 presents preliminary analyses of temporal trends in the therapy received by older women diagnosed with local or regional breast cancer. Women aged 65-74 actually experienced a significant worsening of care. For example, in 1992-95 12.2% of women aged 65-75 with local or regional breast cancer received BCS without radiotherapy compared to 7.8% in 1983-1985. Those differences remain in multivariate analyses controlling for stage, size of tumor and patient characteristics (data not shown).

Table 2. Percentage of Women Aged 65-74 Years with Local or Regional Stage Breast Cancer Who Receive BCS Without Radiotherapy, and the Percent Who Receive Neither Axillary Dissection Nor Radiotherapy, 1983 to 1995, by SEER Area.

SEER Areas	Number of Patients	% of Total Receiving BCS Without Radiation Therapy	% of Total Receiving Neither Radiation Nor Axillary Dissection
San Francisco-Oakland	7,160	9.1	5.3
Connecticut	6,942	16.0	9.6
Metropolitan Detroit	7,665	10.8	6.9
Hawaii	1,657	9.1	5.4
Iowa	5,967	5.2	3.8
New Mexico	2,037	8.1	6.0
Seattle	6,458	8.3	5.5
Utah	2,103	8.4	5.0
Metropolitan Atlanta	2,570	8.4	5.1

Table 3. Percentage of Older Women with Local or Regional Stage Breast Cancer Who Received BCS Without Radiotherapy or Who Received Neither Axillary Dissection Nor Radiotherapy, by Age and Time Period, 1983 to 1995.

Year	Number of Patients	% Receiving BCS without Radiotherapy		% Receiving Neither Axillary Dissection nor Radiotherapy	
		65-74 Years	≥ 75 Years	65-74 Years	> 75 Years
1983-85	14,002	7.8	25.7	5.4	21.8
1986-88	18,011	7.4	23.9	4.5	20.2
1989-91	19,179	9.9	26.1	6.5	22.8
1992-95	27,001	12.2	28.1	7.4	23.9

7. KEY RESEARCH ACCOMPLISHMENTS

- Demonstrated inutility of Medicare data alone to provide valid estimates of breast cancer incidence.
- Demonstration of marked variation in breast cancer mortality rates by Health Service Area.
- Demonstration that variation in breast cancer mortality is correlated with variation by HSA in survival with breast cancer.

8. REPORTABLE OUTCOMES

Manuscripts

Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. J Clin Epidemiol, in press.

Du XL, Freeman JL, Warren JL, Nattinger AB, Zhang D, Goodwin JS. Accuracy and completeness of Medicare claims data for surgical treatment of breast cancer. Medical Care, in press.

Abstracts

None

Funding Applied for and Funded Based on Work Supported by This Award

PI: Xianglin Du (07/01/99-06/30-01)
U.S. Department of Defense, Impact of Axillary Dissection on Clinical Outcomes of Breast Cancer Surgery.

PI: James S. Goodwin (06/01/99-05/31/03)
NIH/NCI, Variations in Breast Cancer Treatment and Mortality

9. CONCLUSIONS

Survival from breast cancer does vary by geographic area.

A major contributor to variation in survival is variation in treatment, in particular, women who receive breast conserving surgery (BCS) without radiotherapy. Most of these women are also not undergoing axillary node dissection, which means that many are not correctly staged. The percentage of older women who receive what is in essence an excisional biopsy without axillary dissection and without radiotherapy has actually increased over time, from 1983 to 1995. In addition, there are a large geographic variations in the percentages of older women receiving this minimal therapy. This increase in the use of BCS without radiotherapy and without axillary dissection is responsible in part of the lack of improvement in overall breast cancer mortality in older women over the past 20 years, while there have been rather impressive reductions in breast cancer mortality rates in younger women (71). More relevant to this proposal, the marked geographic variations in use of BCS without radiotherapy and without axillary dissection lead to geographic variations in survival which in turn contribute to the known geographic variations in breast cancer mortality. It is important to understand that breast cancer patients are primarily treated at community hospitals by general surgeons who perform few such operations yearly. Half of all hospitals performing breast cancer surgery perform 15 or fewer cases per year (60). In other words, older women with breast cancer are not exposed to programs or specialized centers that would ensure compliance with the relatively complex therapy of five time weekly radiotherapy for six weeks.

The major implication of our findings is the need for better standardization of care for the older women with breast cancer.

10. REFERENCES

1. Makuc DM, Halund B, Ingram DD, et al. Health Service Areas for the United States. National Center for Health Statistics. Vital Health Stat. 2(112). 1991.

11. APPENDICES

Copy of manuscript by J.L. Freeman, et al.

LIST OF PERSONNEL RECEIVING PAY FROM THIS RESEARCH EFFORT

James S. Goodwin, M.D.
Jean Freeman, Ph.D.
Daniel Freeman, Ph.D.
Dorothy Syblik, M.S.
Xianglin Du, M.D., Ph.D.

An Approach to Identifying Incident Breast Cancer Cases Using Medicare Claims Data

Jean L. Freeman[1,2,3], Dong Zhang[2], Daniel H. Freeman, Jr.[2], James S. Goodwin[1,3]

[1]Division of Geriatric Medicine
Department of Internal Medicine
University of Texas Medical Branch
Galveston, TX

[2]Division of Epidemiology and Biostatistics
Department of Preventive Medicine & Community Health
University of Texas Medical Branch
Galveston, TX

[3]Sealy Center on Aging
University of Texas Medical Branch
Galveston, TX

An Approach to Identifying Incident Breast Cancer Cases Using Medicare Claims Data

Jean L. Freeman, Dong Zhang, Daniel H. Freeman, Jr., James S. Goodwin

ABSTRACT

This study developed and evaluated a method for ascertaining a newly diagnosed breast cancer case using multiple sources of data from the Medicare claims system. Predictors of an incident case were operationally defined as codes for breast cancer related diagnoses and procedures from hospital inpatient, hospital outpatient and physician claims. The optimal combination of predictors was then determined from a logistic regression model using 1992 data from the linked SEER registries-Medicare claims data base and a sample of non-cancer controls drawn from the SEER areas. While the ROC curve demonstrates that the model can produce levels of sensitivity and specificity above 90%, the positive predictive value is comparatively low (67-70%). This low predictive value is largely the result of the model's limitation in distinguishing recurrent and secondary malignancies from incident cases and possibly from the model identifying true incident cases not identified by SEER. Nevertheless, the logistic regression approach is a useful method for ascertaining incident cases since it allows for greater flexibility in changing the performance characteristics by selecting different cut-points depending on the application (e.g., high sensitivity for registry validation, high specificity for outcomes research). It also allows us to make specific adjustments to population based estimates of breast cancer incidence with claims.

Keywords: breast neoplasms; Medicare; sensitivity and specificity; incidence; registries
Running Title: Identifying Incident Breast Cancer with Medicare Data

INTRODUCTION

Administrative data bases are widely recognized as important sources of data for epidemiologic studies and health services research [1-17]. A major challenge facing investigators utilizing these data is how to identify incident cases of disease. Approaches to identify these newly diagnosed cases have all employed complex algorithms based on combinations of data elements in the files [18-23].

For breast cancer in particular, previous approaches to case ascertainment with Medicare claims, hospital discharge data bases and similar types of administrative data have used combinations of diagnosis and procedure codes from hospital stay or billing records to estimate incidence [20-23], describe patterns of surgical treatment [24-31] and estimate costs of cancer related services [32-39]. The validity of these approaches is supported by further evidence that 1) cancer is a diagnosis reliably coded on hospital billing records [40] and 2) algorithms to estimate cancer incidence rates from hospital claims have produced estimated rates similar to those of cancer registries [20-21]. However, attempts to identify specific cases of breast cancer (as distinguished from overall incidence rates) using administrative data have raised concerns about the **claims'** ability to identify individual subjects with incident disease, particularly with hospital data alone [23,41].

The objective of this study is to develop and evaluate a method for ascertaining a newly diagnosed breast cancer case using multiple sources of diagnoses and procedures from the Medicare claims system. Predictors of an incident case are operationally defined as codes for

breast cancer related diagnoses and procedures. The optimal combination of predictors is then determined from a logistic regression analysis that includes predictors from the year of diagnosis as well as two years prior to diagnosis to eliminate prevalent cases.

MATERIALS AND METHODS

Data sources

We developed our algorithm for ascertaining an incident breast cancer case with Medicare claims data from a sample of women who were confirmed with newly diagnosed breast cancer (cases) and a sample of women without breast cancer (controls). Data on the cases and their claims were obtained from a data base of linked tumor registry records from the Surveillance, Epidemiology, and End Results (SEER) Program and Medicare claims data from the Health Care Financing Administration (HCFA) [13,42]. The SEER Program supports population based tumor registries, which for this study included nine registries from the metropolitan areas of San Francisco/Oakland, Detroit, Atlanta and Seattle and the states of Connecticut, Iowa, New Mexico, Utah and Hawaii [43].

The SEER/Medicare data base was developed as part of a collaborative project between the National Cancer Institute and HCFA [13]. It currently contains data on persons 65 years and older who were diagnosed with any cancer, except non-melanoma skin cancer, from 1973 through 1993 in one of the SEER areas and who linked to the Medicare enrollment file. Records were linked using a deterministic algorithm that declared two records a match if they agreed on

selected combinations of personal identifiers (Social Security Number, first name, last name, middle initial, year of birth, month of birth, day of birth, date of death). Records were successfully linked for 93.8 percent of the SEER registry cases diagnosed at age 65 or older. In addition, the project has drawn five percent samples of non-cancer controls from the nine SEER areas each year using the Medicare enrollment file. These samples were drawn from the enrollment file after excluding all beneficiaries who ever linked to the SEER registries.

For each SEER case, the data base contains information on demographic characteristics, diagnosis date, date of death, tumor characteristics and treatment provided within four months of the initial therapy after diagnosis. For both cases and controls the data base also contains information from the Medicare enrollment file on demographic characteristics, date of death, entitlement, HMO membership and coverage under Parts A (hospital care) and B (physician and outpatient services).

Claims data for the cancer cases and non-cancer controls were extracted from three HCFA files over the period 1990-1992: 1) the Medicare Provider Analysis and Review (MEDPAR) file, 2) the Hospital Outpatient Standard Analysis file (SAF) and 3) the National Claims History File (NCH) [44]. MEDPAR contains claims for hospital inpatient stays covered under Part A. Diagnoses (up to 10) and procedures (up to 10) are coded in the International Classification of Diseases, Clinical Modification (ICD-9-CM) [45]. SAF data contain claims on facility-based outpatient services. Diagnoses and procedures are coded in ICD-9-CM. Procedures are also coded in the HCFA Common Procedure Coding System (HCPCS), which includes Common Procedure Terminology (CPT) codes [46] and other codes assigned by the HCFA local carriers.

The NCH file contains the claims for physicians' and other medical services covered under Part B. Diagnoses are coded in ICD-9-CM and procedures are coded in HCPCS.

In the following description of the methods and results, claims from these files are labeled "hospital inpatient" and "hospital outpatient" if they were found in the MEDAR and SAF files, respectively. They are labeled "physician" if they were found in the National Claims History file.

Study subjects

Cases. Incident breast cancer cases are all females in the SEER tumor registry that were diagnosed with breast cancer at age 65 through 74 in 1992 and linked with the Medicare data. Excluded are women who were members of an HMO at any time during 1992 or who were not covered under both Parts A and B of Medicare for any part of that year. These subjects were excluded since their claims for certain services may not be included in the HCFA data base. There were a total of 4,326 women age 65-74 from the nine SEER registries included in our study. Of these 448 were excluded because of HMO membership. An additional 539 subjects were excluded for lack of Part A and B coverage, resulting in 3,339 eligible cases.

Controls. Control subjects are all women in the five percent non-linked file from the SEER areas who were age 65-74 as of January 1992. As with the cases, control subjects were covered by both Parts A and B in 1992 and not members of an HMO at any time during that year. After these exclusions there were 44,221 control subjects.

Study variables

Information in the claims data that could be used to identify an incident breast cancer case are diagnoses of breast cancer as well as specific services provided to women in the detection, diagnosis and first course of therapy for breast cancer. A list of potential predictors was developed that contained these diagnoses and services. Services were included on the list if they were used in other studies that investigated the use of Medicare claims to identify incident breast cancer care [20-24]. A co-author (JSG) also reviewed the procedure coding manuals and selected additional procedures that he felt were associated with breast cancer diagnosis and treatment.

These predictor variables were then operationally defined in terms of diagnoses coded in ICD-9-CM and procedures coded in ICD-9-CM, CPT and HCPCS. A list of the variables and their corresponding codes appears in Table 1. Breast cancer diagnosis codes in 1992 data are used to identify incident cases of breast cancer and those in 1990-1991 data are used to identify (and eliminate) prevalent cases. The code V103 (personal history of breast cancer) is used only to identify the prevalent cases in 1990-1991.

Development of prediction model

The analysis file contained one record per subject with the outcome variable set to 1 if the subject was a case and set to 0 if she was a control. Dichotomous representations of the predictor variables were generated from each subject's claims with the outcome indicating whether the

particular diagnosis or service was present on a hospital inpatient, physician or hospital outpatient claim for the years 1990, 1991 and 1992.

The outcome was modeled using logistic regression, which can be represented as:

$$\ln(p/(1-p)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \qquad (1)$$

where $p/(1-p)$ is the odds of being a case versus a control, the X_i are variables representing the binary predictors and the β_i are parameters that determine the strength and direction of the associations. Estimates of the β_i are obtained by the method of maximum likelihood. Computations used the Statistical Analysis System (SAS) and its logistic regression procedure, LOGIST [47].

Four logistic regression models were estimated with the claims data. Model 1 (6 predictors) contained only the breast cancer diagnosis predictor variables from the hospital inpatient claims (breast cancer as a principal diagnosis on an inpatient claim in 1992, 1991 or 1990; breast cancer as an additional diagnosis on an inpatient claim in 1992, 1991 or 1990). Model 2 (10 predictors) contained all the variables in Model 1 plus the predictor variables representing breast cancer related procedures on the hospital inpatient claims (mastectomy, partial mastectomy, excisional biopsy, incisional biopsy on an inpatient claim in 1992). Model 3 (36 predictors) contained all the breast cancer predictors from the hospital inpatient, hospital outpatient and physician claims (breast cancer as a principal or first listed diagnosis in 1991 or 1992 on a hospital inpatient, hospital outpatient or physician claim; breast cancer as a principal or first listed diagnosis in 1990 on a hospital inpatient or hospital outpatient claim; breast cancer as an additional diagnosis

in 1991 or 1992 on a hospital inpatient, hospital outpatient or physician claim; breast cancer as an additional diagnosis in 1990 on a hospital inpatient or hospital outpatient claim; mastectomy on a hospital inpatient claim in 1992; partial mastectomy, biopsy, incisional biopsy on any claim in 1992; needle biopsy, chemotherapy on a physician claim in 1992; mammography, other breast cancer related radiology, radiation oncology, laboratory test on a hospital outpatient or physician claim in 1992).

Model 4 was derived from a backwards elimination process that involved including all variables in Model 3, then deleting them one at a time based on the values of their corresponding partial F statistic. The final step produced a model with all terms significant at the 0.05 level.

Evaluation of models

The probability of a subject being a SEER breast cancer case or not was estimated from each of these models using the following formula, which is based on equation (1):

$$p_i = \exp(y_i)/(1+\exp(y_i)) \qquad (2)$$

where

p_i= the probability the i'th subject is a SEER breast cancer case,

$y_i = B_0 + B_1 x_{i1} + \ldots + B_k x_{ik}$,

x_{ij}=the value of the j'th predictor variable (1 or 0) for the i'th subject, and

B_j= the j'th parameter from the logistic regression model.

The sensitivity and specificity of each model was then computed using SEER as the "gold

standard" and different cut points for p_k as alternative criteria for defining a breast cancer case with the claims data. That is, for each p_k we identify all the SEER linked cases (n_{k1}) and all the non-cancer controls (n_{k2}) with a probability greater than or equal to p_k of being a breast cancer case based on their Medicare claims data and the logistic model. The sensitivity of the model is then the percent of all SEER linked cases identified as a breast cancer case by the model (n_{k1}/3,339) x 100 and the false positive rate is the percent of all non-cancer controls identified as a breast cancer case (n_{k2}/44,221) x 100. Specificity is computed as 100 minus the false positive rate.

Receiver Operating Characteristic (ROC) curves were generated for each model and their performance assessed by comparing the areas under the curves (AUC). The AUC can also be rewritten as Somer's *D* where *D=2 x (AUC-0.5)*. We used Somer's *D* since it varies between zero (no information) and one (exact predictor) and can be interpreted as a correlation coefficient.

The relative tradeoffs in positive predictive value were also examined for different combinations of sensitivity and specificity. The positive predictive value (PPV) is the percent of all subjects (SEER cases and non-cancer controls) correctly identified as a breast cancer case with a given model and cut-point. Estimates of the PPV were derived for each model and cut-point by first inflating the non-cancer control sample by its sampling weight (0.05), then taking the ratio of the number of SEER cases identified as breast cancer cases divided by the sum of that number plus the weighted number of controls identified as breast cancer cases (false positives).

RESULTS

Distribution of predictors

All predictors (Table 1) that were present in the claims for at least five percent of the SEER cases were included in the logistic regression analysis. The distribution of these predictors in the claims for both the SEER and non-cancer subjects for 1992 is given in Table 2. The percent of SEER subjects (in parentheses) with the predictor is a measure of that predictor's sensitivity in identifying cases. Likewise, the specificity of the predictor is 100 minus the corresponding percent of non-cancer cases with the predictor. The likelihood ratio (ratio of sensitivity to 100-specificity) provides some indication of the ability of that predictor to discriminate cases from non-cases (good discriminators have higher likelihood ratios).

Table 2 allows us to make some initial observations about the potential utility of each 1992 predictor in identifying incident cases. Predictors based on hospital inpatient claims appear to be better discriminators of cases versus non-cancer controls than predictors based on outpatient or physician claims. Their likelihood ratio statistics and predictive values are generally higher compared to other predictors. Their sensitivity is low, however, since only about two-thirds of the subjects had a hospital claim with a principal diagnosis of breast cancer. Highest levels of sensitivity were found for predictors based on physician claims with a first listed diagnosis of breast cancer (95.2 percent), physician claims with a lab procedure (86.8 percent) and physician claims for mammography (86.7 percent).

As expected based on the low incidence of breast cancer, specificity is high (97 percent or above) for all 1992 predictors except mammography. However, the specificity was not high enough to attain levels of positive predictive value above 74 percent for the inpatient predictors and 48 percent for all but two (mastectomy, nodal dissection) of the non-inpatient predictors.

Logistic regression

The four logistic regression models assess the roles that different combinations of predictors play in identifying incident breast cancer cases. Table 3 presents each of the models' estimated parameter estimates and associated p-values.

Models 1 and 2 include only information from the hospital stays. Somer's D statistic is similar for the two models (0.71 vs. 0.72), indicating that there is no additional explanatory power generated by adding the breast cancer procedures (mastectomy, partial mastectomy, excisional biopsy, incisional biopsy) to the breast cancer diagnoses when using only MEDPAR claims.

Explanatory power is considerably increased (Somer's $D = 0.98$) by including information on breast cancer diagnoses and services from all three files in Models 3 and 4. Moreover, there is no improvement in explanatory power with the inclusion of all predictors (Model 3) compared to only the significant predictors (Model 4).

The magnitude and direction (positive or negative) of the parameters reflect the relative importance of the variables in detecting an incident case. Based on Model 4, the likelihood of

being an incident breast cancer case in 1992 increases with having: an inpatient, outpatient or physician claim with breast cancer as a principal or first listed diagnosis in 1992; a physician claim with breast cancer as an additional diagnosis in 1991 or 1992; an inpatient claim with breast cancer as an additional diagnosis in 1992; and a physician claim for partial mastectomy, excisional biopsy, incisional biopsy, needle biopsy, breast cancer related radiation oncology, breast cancer related laboratory test in 1992. The likelihood if being an incident case decreases with having: a hospital outpatient or physician claim with breast cancer as a principal or first listed diagnosis in 1991; a hospital outpatient claim with breast cancer as an additional diagnosis in 1991; a hospital inpatient claim with partial mastectomy in 1992; and a hospital outpatient or physician claim for mammography, other breast cancer related radiology procedure in 1992.

ROC Curve

An ROC curve for Model 4 is presented in Figure 1. Since the model has high specificity, Figure 1 shows only that part of the ROC curve above 98.4 percent specificity. Also shown in the figure is the positive predictive value (PPV) for selected levels of specificity. An optimum cut-point was arbitrarily defined as sensitivity and specificity at least 90 percent and positive predictive value at least 70 percent. This was achieved for a specificity of 99.86 percent and a sensitivity of 90 percent. The cut-point corresponds to a predicted probability of 0.822, derived from the parameters of logistic Model 4 with parameter estimates in Table 3.

At this cut-point, the model generated 62 false positives from the control sample. Since the false positive rate drives the positive predictive value, the individual claims for these false positives

were examined over 1992 to identify characteristics of their medical care that might suggest further refinements to the model. Their diagnosis and treatment patterns based on the claims are summarized in Table 4. About half of the false positives (29 out of 62) had the usual course of care for an early stage incident breast cancer case - biopsy followed by mastectomy or partial mastectomy plus radiation. Another 15 subjects had major parts of this therapy: mastectomy or breast conserving treatment (partial mastectomy/excisional biopsy plus radiation). Among the remaining 18 false positives, 15 appeared to be associated with biopsies to rule out breast cancer (no consistent reporting of breast cancer diagnosis in the claims post biopsy) and three had claims with a breast cancer diagnosis in addition to claims for both radiation therapy and chemotherapy.

We also examined the age distribution of the false positives to see if subjects age 65-66 had a higher likelihood of being identified as a false positive. These subjects are less likely to have claims in 1990 and 1991 (before they were eligible for Medicare) and hence less likely to be eliminated as a prevalent case by the model than older women. Of the 62 false positives, 7 (11 percent) were age 65-66. This is actually less than the proportion of women 65-66 in the study (20 percent).

DISCUSSION

The ability of Medicare claims data to identify incident cases of breast cancer is critical to the data's utility for monitoring trends in incidence and survival, assessing outcomes of treatment, describing patterns of care and estimating costs of disease. The development and evaluation of

alternative approaches to case ascertainment is now feasible with the linked SEER-Medicare data base.

We developed an approach for identifying incident breast cancer cases that is derived from a logistic regression model, which contains variables that indicate the presence or absence of breast cancer related diagnoses and procedures in three sources of claims data: hospital inpatient stays, hospital outpatient services and physician services. Variables representing the entire range of breast cancer related services were examined, including surgery, diagnostic radiology, radiation therapy, chemotherapy and pathology. The final model includes only those variables that had a significant effect on distinguishing newly diagnosed breast cancer cases from controls in the SEER-Medicare data base.

The performance of the model was evaluated in terms of its sensitivity, specificity and positive predictive value for different cut points defined in terms of the predicted probabilities of being an incident case. The ROC curve demonstrates that the model can produce high levels (over 90 percent) of sensitivity and specificity, but the positive predictive value is comparatively low. Using a cut point of 0.822, for example, results in a sensitivity of 90 percent, a specificity of 99.86 percent and a positive predictive value of 70 percent. An examination of the false positives, however, indicates that the model may be detecting recurrent or secondary malignancies as well as incident cases. In fact, about 75 percent of the false positives appear to be subjects receiving services (mastectomy, partial mastectomy, radiation therapy, chemotherapy) for treatment of either newly diagnosed primary, secondary or recurrent breast cancer.

Other methods for identifying incident breast cancer have utilized only the diagnosis codes or some combination of the diagnosis and surgical procedure codes from the hospital inpatient files over a specific time period [20-23]. Subjects with a breast cancer diagnosis on a claim before the period were considered prevalent cases and eliminated. Such approaches appeared promising on data collected in the 1980's, since comparisons of claims based incidence rates were similar to those produced with the SEER registry [20-21]. However, when comparisons were made at the individual level between hospital claims and tumor registry data, discrepancies were found in the specific cases identified as incident breast cancer by the two sources [23], suggesting that "compensating errors" may have given rise to the similar population based estimates of incidence rates [23]. Moreover, a 1989 comparison between SEER and Medicare data found that case ascertainment through hospital claims had only a sensitivity of 59.4 percent and a specificity of 96.6 percent [22].

We also found that use of hospital inpatient data alone produces significantly lower sensitivity for given levels of specificity than use of three data files. This is most likely attributed to the shift in surgery from the inpatient to the outpatient setting, particularly since 1990. In fact, only two-thirds of the SEER breast cancer subjects in 1992 had any hospital inpatient claim.

The higher levels of sensitivity and specificity achieved with the logistic regression model can be attributed not only to the use of claims files from both the inpatient and outpatient settings, but also to the implicit "weighting" of information in these claims as defined by the model's parameters. Alternative methods for identifying incident breast cancer cases are based on combinations of diagnosis and procedure codes that are clinically meaningful but do not

necessarily represent the optimal mix of data that can identify an incident case. Based on the considerable number of plausible diagnosis and treatment patterns for incident breast cancer, it would be extremely difficult for clinicians a priori to construct a "decision tree" with associated probabilities that defines the criteria for identifying a newly diagnosed breast cancer case with the claims data. The logistic regression model accomplishes this through statistical algorithms that find the optimal mix of variables that distinguishes women with incident breast cancer from other women.

While the logistic regression model can generate higher levels of sensitivity and specificity compared to alternative approaches based on hospital claims, it has several limitations. First, the model is not immediately interpretable or medically meaningful to anyone, let alone clinicians. It is a formula that generates the estimated probabilities of a subject being an incident breast cancer case based on the services utilized over a period of time. In fact, the optimal mix of variables as defined by the model's parameters produces unexpected results, such as an increased likelihood of being an incident case with an additional diagnosis of breast cancer on a physician claim in 1991. However, the face validity of the model is enhanced by examining the relative probabilities it generates for different combinations of services.

Second, the model may not be generalizable across different time periods. The increasing trend toward more outpatient services, that is also likely to include more extensive diagnostic work-up and adjuvant therapy for older women over time, could change the relative weights of claims information. The stability of the model's parameter estimates across time periods will be an important extension of this research.

Third, due to the low incidence of breast cancer, high levels of positive predictive value can only be achieved with extremely high rates of specificity. Specificity is increased by decreasing the number of false positives. An examination of the model's false positives highlights the challenges of using claims data to identify incident cases of breast cancer and the limitations in using the linked SEER Medicare data base as a gold standard for assessing the model's performance. Almost three quarters of the false positives had patterns of medical care in the claims that were the same as those for a typical incident case. These false positives may be recurrent disease for subjects initially diagnosed outside the SEER areas, incident cases missed by SEER or cases that SEER identified but were not linked to the Medicare enrollment files. In fact, in a study using Medicare inpatient data, Warren et. al. [22] also found false positive cases with a breast cancer diagnosis and a mastectomy procedure code. These cases were sent to the SEER registries for follow-up and a number were found to be incident cases that did not link with the Medicare data.

It is also important to note that new breast cancer diagnosed in the contralateral breast is considered an incident case by SEER. Hence, any of these cases identified by the algorithm are in fact "true" positives. Of the 3,339 cases, 218 were previously reported by the SEER registries and 189 of the 218 were identified by our algorithm. Such cases could never be identified by the claims data alone since laterality cannot be determined with ICD-9-CM diagnosis (breast cancer) or procedure (mastectomy) codes.

A fourth limitation is that the non-cancer control sample was drawn from the population of Medicare beneficiaries in the SEER areas who never linked with any case in the SEER registry

since 1973. Hence, the control sample contains no subjects who were previously diagnosed with cancer by SEER (i.e., a prevalent case). Not including them in the controls inflates our estimates of specificity and positive predictive value. We estimated the extent of this inflation by applying our algorithm to the claims for all breast cancer cases diagnosed from 1984 through 1991, who would have been alive and age 65 to 74 in 1992. Using the same cut-point (p=0.822), the algorithm identified 212 of these prevalent cases as incident. If we assume all these cases still lived in the SEER area, their false positive status would reduce our positive predictive value from 70 to 67 percent.

Limitations of generalizability and test performance (sensitivity, specificity, positive predictive value) are common to any approach that utilizes the claims data to identify incident disease. Validation studies using gold standards such as SEER can produce these performance measures, which allow us to make specific adjustments to our population based estimates of breast cancer incidence using a claims based algorithm.

A particular advantage of using a logistic model to distinguish cases from non-cases is that it allows for greater flexibility in changing the performance characteristics by selecting different cut-points, depending on the application. For example, we may want to conduct a survey and medical chart review of older women with breast cancer in a region without a tumor registry. A cut-point with high sensitivity can be used to capture 95 percent of all the cases. While only about 50 percent (positive predictive value) will be true cases, the false positives can be eliminated through questions that "screen out" ineligible women prior to conducting the interview. Likewise, in regions with a cancer registry, this approach can also be used to assess

the completeness of the registry by examining the medical records of persons identified with Medicare but not in the registry. Other applications, such as estimating outcomes of care, require high positive predictive value and a cut point can be selected that identifies a high proportion of true incident cases.

In conclusion, the logistic regression approach to identifying incident breast cancer with three sources of claims (hospital inpatient, hospital outpatient, physician services) produces superior estimates of sensitivity, specificity and positive predictive value compared to alternative approaches based on hospital data alone or specific sets of diagnosis and procedure codes. While the algorithm produces levels of sensitivity and specificity above 90 percent, the positive predictive value at these levels is comparatively low (67 to 70 percent). This low predictive value is largely the result of the model's limitations in distinguishing recurrent and secondary malignancies from incident cases and possibly from the model identifying true incident cases not identified by SEER. Nevertheless, the logistic regression approach is a useful method for ascertaining incident cases since it allows for greater flexibility in changing the performance characteristics by selecting different cut-points depending on the application (e.g., high sensitivity for registry validation, high specificity for outcomes research). It also allows us to make specific adjustments to population based estimates of breast cancer incidence with claims.

ACKNOWLEDGMENT

This research was supported by grants from the National Cancer Institute (CA72076) and the Department of Defense (DAMD17-97-1-709-5). The study used the Linked SEER-Medicare Data Base. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the Applied Research Branch, Division of Cancer Prevention and Population Science, NCI; the Office of Information Services, and the Office of Strategic Planning, HCFA; Information Management services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database.

Figure 1. Partial ROC curve and positive predictive value of logistic regression model for ascertaining incident breast cancer with Medicare claims

Table 1. List of breast cancer predictors based on diagnoses and procedures that are defined with claims data using ICD-9-CM diagnosis and procedure codes, CPT and HCPCS procedure codes, and revenue center codes

Breast Cancer Predictors	ICD-9-CM Codes	CPT/HCPCS Codes	Revenue Center Codes
Breast Cancer Dx.			
1992*	174xx, 2330		
1990-1991*	174xx, 2330, V103		
Mastectomy	8541-8549	19180, 19200-19220,19240	
Partial Mastectomy	8522, 8523	19160, 19162	
Excisional Biopsy	8521	19120	
Axillary Nodal Dissection	403	38740, 38745	
Needle Biopsy	8511	19100	
Incisional Biopsy	8512	19101	
Cyst Aspiration		19000, 19001	
Radiologic Diagnostic Procedures:			
Mammography	8737	76090-76092	
Other†	8735, 8736, 8873, 8885	76003, 76645, 76086-76088, 76095-76365	
Bone Scan		76061, 76062	
Radiation Oncology	9221-9229, 9985	77261-77499, 77600-77620, 77750-77799, 79200-79999	330, 333, 339, 342
Chemotherapy	9925	96400-96549, J9000-J9999	331, 332, 335
Laboratory‡		88307, 88309, 84233, 84234 88329, 88331, 88332	

Note: ICD-9-CM codes are used to describe diagnoses and procedures according to the International Classification of Diseases, Ninth Revision, Clinical Modification. CPT and HCPCS codes describe physician services and procedures according to the Common Procedure Terminology (CPT) and the HCFA Common Procedure Coding System (HCPCS), respectively.

* Breast cancer diagnosis codes in 1992 data are used to identify incident cases of breast cancer and those in 1990-1991 data are used to identify (and eliminate) prevalent cases. The code V103 (personal history of breast cancer) is used only to identify the prevalent cases in 1990-1991.

† For example, localization for breast biopsy, breast ultrasound, breast thermography.

‡ For example, surgical pathology and estrogen receptor assay.

Table 2. Number (percent) of SEER breast cancer cases and non-cancer controls that have breast cancer predictors in claims, the likelihood ratio and the positive predictive value of being an incident case: distribution according to breast cancer predictor and type of claim

Breast Cancer Predictor and Type of Claim	SEER	Non-Cancer	Likelihood Ratio	Positive Predictive Value*
Breast Cancer Principal/ First Listed Dx				
hospital inpatient	2266 (67.9)	40 (.1)	750.3	73.9
hospital outpatient	2017 (60.4)	413 (.9)	64.7	19.6
physician	3181 (95.2)	962 (2.2)	43.8	14.2
Breast Cancer Other Dx				
hospital inpatient	208 (6.2)	22 (.0)	125.2	32.1
hospital outpatient	304 (9.1)	65 (.1)	61.9	19.0
physician	1913 (57.3)	273 (.6)	92.8	25.9
Mastectomy				
hospital inpatient	1776 (53.2)	33 (.1)	712.8	72.9
physician	1771 (53.0)	35 (.1)	670.1	71.2
Partial Mastectomy				
hospital inpatient	246 (7.4)	7 (.0)	465.4	63.7
hospital outpatient	250 (7.5)	19 (.0)	174.3	40.0
physician	921 (27.6)	49 (.1)	248.9	48.4
Excisional Biopsy				
hospital inpatient	284 (8.5)	11 (.0)	341.9	56.3
hospital outpatient	1171 (35.1)	215 (.5)	72.1	21.4
physician	1869 (56.0)	299 (.7)	82.8	23.8
Nodal Dissection				
physician	312 (9.3)	2 (.0)	2066.0	88.6
Needle Biopsy				
Physician	407 (12.2)	60 (.1)	89.8	25.3

Incisional Biopsy				
hospital inpatient	266 (8.0)	5 (.0)	704.6	72.7
hospital outpatient	431 (12.9)	84 (.2)	68.0	20.4
physician	291 (8.7)	46 (.1)	83.8	24.0
Mammography				
hospital outpatient	1291 (38.7)	6587 (14.9)	2.6	0.1
physician	2896 (86.7)	14112 (31.9)	2.7	0.1
Other Radiologic Procedures				
hospital outpatient	1010 (30.3)	569 (2.5)	23.5	8.2
physician	1735 (52.0)	1049 (2.4)	21.9	7.6
Radiation Oncology				
hospital outpatient	703 (21.1)	106 (.2)	87.8	24.9
physician	1011 (30.3)	160 (.4)	83.7	24.0
Chemotherapy				
physician	363 (10.9)	356 (.8)	13.5	4.9
Laboratory				
outpatient	1243 (37.2)	311 (.7)	52.9	16.6
physician	2898 (86.8)	850 (1.9)	45.2	14.6

---

* The positive predictive value is based on the weighted number of non-cancer controls.

Table 3. Estimated logistic regression parameters (B) and standard errors (S.E.) for alternative models containing breast cancer predictors

Factors	Model 1		Model 2		Model 3		Model 4	
	B	S.E.	B	S.E.	B	S.E.	B	S.E.
Intercept	-3.83	0.03***	-3.84	0.03***	-5.99	0.10***	-5.92	0.09***
Breast Ca. Principal/ First Listed Dx.								
hospital inpatient 1992	7.86	0.16***	6.04	0.26***	3.00	0.58***	4.13	0.27***
hospital inpatient 1991	0.82	0.63	0.84	0.62	-0.69	0.74		
hospital inpatient 1990	-0.26	0.91	-0.50	0.89	-1.37	0.79		
hospital outpatient 1992					1.76	0.21***	1.71	0.19***
hospital outpatient 1991					-1.61	0.39***	-1.76	0.36**
hospital outpatient 1990					-0.47	0.39		
physician 1992					3.69	0.16***	3.78	0.16***
physician 1991					-2.31	0.25***	-2.46	0.24***
Breast Ca. Other Dx.								
hospital inpatient 1992	5.68	0.25***	5.30	0.26***	1.58	0.47**	1.78	0.41***
hospital inpatient 1991	-1.86	0.78*	-1.71	0.79*	-1.37	0.88		
hospital inpatient 1990	-2.67	0.77**	-2.32	0.81**	-2.28	1.24		
hospital outpatient 1992					0.57	0.38		
hospital outpatient 1991					-2.25	0.62**	-2.45	0.59***
hospital outpatient 1990					0.22	0.65		
physician 1992					1.37	0.19***	1.44	0.18***
physician 1991					1.23	0.27***	1.03	0.26***
Mastectomy								
hospital inpatient 1992			2.00	0.31***	0.92	0.59		
Partial Mastectomy								
hospital inpatient 1992			2.06	0.55**	-1.61	0.71*	-2.23	0.59***
hospital outpatient 1992					-0.08	0.47		
physician 1992					2.22	0.32***	2.28	0.28***
Excisional Biopsy								
hospital inpatient 1992			3.03	0.48***	0.95	0.57		
hospital outpatient 1992					-.01	0.27		
physician 1992					2.07	0.22***	2.08	0.20***

Needle Biopsy							
physician 1992				2.36	0.34***	2.35	0.34***
Incisional Biopsy							
hospital inpatient 1992		0.62	0.57	1.12	0.66		
hospital outpatient 1992				0.17	0.30		
physician 1992				1.90	0.34***	1.93	0.33***
Mammography							
hospital outpatient				-0.62	0.18**	-0.50	0.16*
physician				0.25	0.15		
Other Rad. Proc.							
hospital outpatient 1992				-0.55	0.26*	-0.69	0.24*
physician 1992				1.19	0.20***		
Radiation Oncology							
hospital outpatient 1992				-0.30	0.41		
physician 1992				0.69	0.35*	1.29	0.20***
Chemotherapy							
physician 1992				0.04	0.31		
Laboratory							
outpatient 1992				-0.30	0.26		
physician 1992				0.76	0.19***	0.80	0.18***
Somer's D	0.71		0.72	0.98		0.98	

Models 1: Breast cancer diagnosis (principal or additional diagnosis) on a 1990, 1991, or 1992 hospital inpatient claim

2: Breast cancer diagnosis (principal or additional diagnosis) on a 1990, 1991 or 1992 hospital inpatient claim or breast cancer related procedure (mastectomy, partial mastectomy, excisional biopsy, incisional biopsy) on a 1992 hospital inpatient claim

3: All breast cancer related predictors from all files: breast cancer as a principal or first listed diagnosis in 1991 or 1992 on a hospital inpatient, hospital outpatient or physician claim; breast cancer as a principal or first listed diagnosis in 1990 on a hospital inpatient or hospital outpatient claim; breast cancer as an additional diagnosis in 1991 or 1992 on a hospital

inpatient, hospital outpatient or physician claim; breast cancer as an additional diagnosis in 1990 on a hospital inpatient or hospital outpatient claim; mastectomy on a hospital inpatient claim in 1992; partial mastectomy, biopsy, incisional biopsy on any claim in 1992; needle biopsy, chemotherapy on a physician claim in 1992; mammography, other breast cancer related radiology, radiation oncology, laboratory test on a hospital outpatient or physician claim in 1992

4: Significant ($p<0.05$) breast cancer related predictors from all files: breast cancer as a principal or first listed diagnosis in 1992 on a hospital inpatient, hospital outpatient or physician claim; breast cancer as a principal or first listed diagnosis in 1991 on a hospital outpatient or physician claim; breast cancer as an additional diagnosis in 1992 on a hospital inpatient or physician claim; breast cancer as an additional diagnosis in 1991 on a hospital outpatient or physician claim; partial mastectomy in 1992 on a hospital inpatient or physician claim; excisional biopsy, needle biopsy, incisional biopsy, radiation oncology, laboratory test in 1992 on a physician claim; mammography, other radiological procedures in 1992 on a hospital outpatient claim

* p-value <.05
** p-value <.01
*** p-value < .0001

Table 4. Distribution of false positive control subjects by diagnosis and breast cancer related treatment patterns in the 1992 claims data

Pattern	Number of False Positives
Breast cancer dx +biopsy + mastectomy	22
Breast cancer dx +biopsy + partial mastectomy + radiation	7
Breast cancer dx +mastectomy	7
Breast cancer dx + partial mastectomy or excisional biopsy + radiation	8
Breast cancer dx + chemotherapy + radiotherapy	3
Probably rule out biopsy	15
	62

REFERENCES

[1] Roos LL, Nicol JP, Cageorge SM. Using administrative data for longitudinal research: comparisons with primary data collection. J Chronic Dis 1987; 40: 41-49.

[2] Wennberg JE, Roos N, Sola L, et. al. Use of claims data systems to evaluate health care outcomes. Mortality and re-operation following prostatectomy. JAMA 1987;257:933-936.

[3] Connell FA, Diehr P, Hart LG. The use of large data bases in health care studies. Annu Rev Public Health 1987;8:51-74.

[4] Bright RA, Avorn J, Everitt DE. Medicaid data as a resource for epidemiologic studies: strengths and limitations. J Clin Epidemiol 1989; 42: 937-945.

[5] Ray WA, Griffin MR. Use of Medicaid data for pharmacoepidemiology. Am J Epidemiol 1989; 129: 837-849.

[6] Anderson G, Steinberg EP, Whittle J, et. al. Development of clinical and economic prognoses from Medicare claims data. JAMA 1990; 263: 967-972.

[7] Gerstman BB, Lundin FE, Stadel BV, et. Al. A method of pharmacoepidemiologic analysis that uses computerized Medicaid. J Clin Epidemiol 1990; 43: 1387-1393.

[8] Ku LM, Ellwood MR, Klemm J. Deciphering Medicaid data: issues and needs. Health Care Financing Rev 1990; Annual Supplement : 35-45.

[9] Epstein MH. Guest alliance: uses of state-level hospital discharge data bases. J AHIMA 1992; 63: 32-37.

[10] Hannan EL, Kilburn H, et. al. Clinical versus administrative data bases for CABG surgery: does it matter? Med Care 1992; 30: 892-907.

[11] Quam L, Ellis BM, Venus P, et. Al. Using claims data for epidemiologic research: the concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population. Med Care 1993; 31: 498-507.

[12] Romano PS. Can administrative data be used to compare the quality of health care? Med Care Rev 1993; 50: 451-477.

[13] Potosky AL, Riley GF, Lubitz JD, et. al. Potential for cancer related health services research using a linked Medicare-Tumor Registry data base. Med Care 1993;31:732-747.

[14] Romano PS, Roos LL, Luft HS, et. al. A comparison of administrative versus clinical data: coronary artery bypass surgery as an example. J Clin Epidemiol 1994; 47: 249-260.

[15] Lave JR, Pashos CL, Anderson GF, et. al. Costing medical care: Using Medicare

administrative data. Med Care 1994; 32: JS77-JS89.

[16] Fowles JB, Lawthers JP, et. Al. Agreement between physicians' office records and Medicare Part B Claims data. Health Care Financing Rev 1995; 16 (4): 189-199.

[17] Garnick DW, Hendricks AM, Comstock CV, et. Al. A guide to using administrative data for medical effectiveness research. J Outcomes Manage 1996; 3: 18-23.

[18] Baron JA, Lu-Yao G, Barrett J, et. al. Internal validation of Medicare claims data. Epidemiology 1994;5:541-544.

[19] Ray WA, Griffin MR, Fought RL, et. al. Identification of fractures from computerized Medicare files. J Clin Epidemiol 1992;45:703-714.

[20] Whittle J, Steinberg EP, Anderson GF, et. al. Accuracy of Medicare claims data for estimation of cancer incidence and resection rates among elderly Americans. Med Care 1991;29:1226-1236.

[21] McBean AM, Warren JL, Babish JD. Measuring the incidence of cancer in elderly Americans using Medicare claims data. Cancer 1994; 73:2417-2425.

[22] Warren JL, Riley GF, McBean AM, et. al. Use of Medicare data to identify incident breast cancer patients. Health Care Financing Rev 1996 (18):237-246.

[23] McClish DK, Penberthy L, Whittemore M, et. al. Ability of Medicare claims data and cancer regisries to identify cancer cases and treatment. Am J Epidemiol 1997; 145: 227-233.

[24] Nattinger AB, Gottlieb MS, Veum J, et. al. Geographic variation in the use of breast-conserving treatment for breast cancer. N Eng J Med 1992;326:1102-1107.

[25] Ballard-Babash R, Potosky AL, Harlan LC, et. al. Factors associated with surgical and radiation therapy for early stage breast cancer in older women. J Natl Cancer Inst 1996; 88: 716-726.

[26] Nattinger AB, Gottlieb MS, Hoffman RG, et.al. Minimal increase in use of breast-conserving surgery from 1986 to 1990. Med Care 1996; 34(5):479-489.

[27] Du X, Freeman JL, Goodwin JS. The declining use of axillary dissection in patients with early stage breast cancer. Breast Cancer Res Treat (in press).

[28] Hillner BE, Penberthy L, Desch CE, et. al. Variation in staging and treatment of local and regional breast cancer in the elderly. Breast Cancer Res Treat 1996; 40: 75-86.

[29] Kleinman JC, Machlin SR, Madans J, et. al. Changing practice in the surgical treatment of breast cancer. Med Care 1983; 21:1232-1242.

[30] Greenberg ER, Stevens M. Recent trends in breast surgery in the United States and United

Kingdom. BMJ 1986; 292:1487-1491.

[31] Kahn LH, Blustein J, Arons RR, et. al. The validity of hospital administrative data in monitoring variations in breast cancer surgery. Am J Public Health 1996; 86: 243-245.

[32] Riley G, Lubitz J, Prihoda R et. al. The use and costs of medicare services by cause of death. Inquiry 1987; 24:233-244.

[33] Riley GF, Potosky AL, Lubitz JD, et. al. Medicare payments from diagnosis to death for elderly cancer patients by stage at diagnosis. Med Care 1995; 33: 828-841.

[34] Baker MS, Kessler LG, Urban N, et. al. Estimating the treatment costs of breast and lung cancer. Med Care 1991;29:40-49.

[35] Scotto J, Chiazze L. Third National Cancer Survey: Hospitalizations and Payments to Hospitals, Part A: Summary. Washington DC: U.S. Department of Health, Education and Welfare, 1976. (DHEW Pub. No. (NIH) 76-1094).

[36] Iezzoni LI, Henderson MG, Bergman A, et. al. Purpose of admission and resource use during cancer hospitalizations. Health Care Financing Rev 1991; 13:29-40.

[37] Munoz E, Chalfin D, Rosner F, et. al. Hospital costs, cancer patients and medical diagnosis related groups. Oncology 1988; 45:401-404.

[38] Munoz E, Chalfin D, Sterman H, et. al. Hospital costs, cancer patients and surgical diagnostic related groups. J Surg Oncol 1989; 41:47-51.

[39] Litvak S, Borrerro E, Katz R, et. al. Early discharge of the post-mastectomy patient: Unbundling of hospital services to improve profitability under DRGs. Ann Surgery 1987; 53:577-576.

[40] Fisher ES, Whaley FS, Krushat WM, et. al. The accuracy of Medicare's hospital claims data; progress has been made but problems remain. Am J Public Health 1992; 82: 243-248.

[41] Solin LJ, Legorreta A, Schultz DJ, et. al. Analysis of a claims database for identification of patients with carcinoma of the breast. J Med Syst 1994; 18:23-32.

[42] Reynolds T. Linked databases help researchers track costs, care patterns, outcomes. J Natl Cancer Inst 1994; 86: 168-171.

[43] Ries LAG, Kosary CL, Hankey BF, et. al. SEER Cancer Statistics Review, 1973-1994. Bethesda MD: National Cancer Institute, 1997. (NIH Publication No. 97-2789).

[44] Data Users Reference Guide. Health Care Financing Administration. U.S. Department of Health and Human Services, September 1995.

[45] U.S. Public Health Service. International Classification of Diseases, 9th Revision, Clinical

Modification, 5th edition. Los Angeles CA: Practice Management Information Corporation.

[46] American Medical Association. Physicians' Current Procedural Terminology - CPT 94. Chicago: American Medical Association, 1993.

[47] Stokes ME, Davis CE, Koch GG. Categorical Data Analysis Using the SAS System. Cary: SAS Institute Inc, 1997.