Published in final edited form as:

Breast J. 2020 July ; 26(7): 1472–1474. doi:10.1111/tbj.13758.

Using diagnosis codes in claims data to identify cohorts of breast cancer patients following initial treatment

Benjamin L. Franc¹, Robert Thombley¹, Yanting Luo², John Boscardin³, Hope S. Rugo⁴, David Seidenwurm⁵, R. Adams Dudley⁶

¹Department of Radiology, Division of Nuclear Medicine and Molecular Imaging, Stanford University School of Medicine, Stanford, CA, USA.

²Philip R. Lee Institute for Health Policy Studies, University of California San Francisco, San Francisco, CA, USA.

³Department of Medicine, Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA.

⁴Department of Medicine, UCSF Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, USA.

⁵Sutter Medical Foundation, Sacramento, CA, USA.

⁶Philip R. Lee Institute for Health Policy Studies, Center for Healthcare Value, University of California, San Francisco, CA, USA.

Studying patterns of care and medical service utilization after the period of initial treatment of cancer using increasingly-available administrative claims data [1] can be challenging. [2, 3] Patients diagnosed early in their disease may be cured or their cancer may go into remission; therefore, it is possible that few, if any, medical claims remote from the initial treatment will contain a diagnosis code specifically indicating cancer. Patients previously treated for cancer may be overlooked in cancer-related studies of claims data or mistakenly included as a normal control. [4]

We wanted to understand how best to identify breast cancer (BC) patients in claims data during the first 5 years after treatment for the primary cancer. We also aimed to identify any additional diagnosis codes potentially related to the treatment of BC that could be used to identify BC patients whose claims lacked cancer-specific diagnosis codes (174.X family of ICD-9 codes and v10.3).

Using claims (“NCH” and “OUTSAF”) data of 51,278 (median 120 claims) newly diagnosed BC patients (females, age ≥ 65 years, first malignant non-metastatic BC, no non-BC, 6 years of consistent and exclusive enrollment in Medicare parts A and B) from the BC subset of the 2000-2014 Surveillance, Epidemiology and End Results (SEER)-Medicare linked database (N=51,278), we determined the fraction of BC patients who could be identified using BC-specific codes 174.x and v10.3 during the year of diagnosis/treatment

and during each of the 5 years following treatment. Only 551 BC patients lacked a claim containing 174.x or v10.3 diagnosis codes (Table 1).

We then developed a list of diagnosis codes apart from 174.x and v10.3 found more often in claims of BC patients than in claims of a cohort of patients without BC (non-BC cohort of 174,903 patients of a 5% Medicare sample who did not appear in any SEER cancer registry (2000-2014) and did not have claims with 174.x or v10.3 diagnosis codes) using a patient-referenced odds ratio (OR):

$$OR = \frac{f_{BC(no\ 174.x / v10.3),\ ICD-9(XXX.X)}}{f_{nonBC,\ ICD-9(XXX.X)}} = \frac{N_{BC(no\ 174.x / v10.3)\ with\ ICD-9(XXX.X)}}{N_{BC(no\ 174.x / v10.3)\ without\ ICD-9(XXX.X)}} \cdot \frac{f_{nonBC,\ ICD-9(XXX.X)}}{N_{nonBC\ with\ ICD-9(XXX.X)}} = \frac{N_{BC(no\ 174.x / v10.3)\ without\ ICD-9(XXX.X)}}{N_{nonBC\ without\ ICD-9(XXX.X)}}$$

where

$f_{BC(no\ 174.x/v10.3),ICD-9(XXX.X)}$ is the odds that the specific ICD-9 diagnosis code (XXX.X) is present in claims of BC patients who had medical claims but none of those claims contained 174.x/v10.3 codes; $f_{nonBC,ICD-9(XXX.X)}$ is the odds that the specific ICD-9 (XXX.X) is present in the claims of patients without BC;

$N_{BC(no\ 174.x/v10.3)with\ ICD-9(XXX.X)}$ is the number of BC patients with a claim containing the specified ICD-9 code (XXX.X) but no claims containing 174.x/v10.3 codes;

$N_{BC(no\ 174.x/v10.3)without\ ICD-9(XXX.X)}$ is the number of BC patients without a claim containing the specified ICD-9 code (XXX.X) and no claims containing 174.x/v10.3 codes;

$N_{nonBC\ with\ ICD-9(XXX.X)}$ is the number of non-BC patients with a claim containing the specified ICD-9 code (XXX.X); and $N_{nonBC\ without\ ICD-9(XXX.X)}$ is the number of non-BC patients without a claim containing the specified ICD-9 code (XXX.X).

Diagnosis codes over the 5-year follow-up period that appeared in at least 0.1% of the population, provided O.R.>1.01, and were related to the management of BC are provided (Table 2).¹

For each year following the initial year of treatment and overall, BC patients were identified as belonging to one of 5 subgroups (Figure 1):

1. Those with at least one utilization claim containing a diagnosis code indicating active, invasive BC (174.x)
2. Those without a 174.x code on claims, but with at least one utilization claim containing a diagnosis code indicating a personal past history of BC (v10.3)
3. Those without any 174.x or any v10.3 codes on claim, but with at least one utilization claim containing a diagnosis code more likely to be encountered in a claim from a patient with BC (O.R>1.01) and related to management of BC

¹All diagnosis codes with O.R. > 1.01 in at least 0.1% of patients may be found in supplemental material 1

4. Those with claims containing only diagnosis codes unrelated to management of BC
5. Those without any claims

In any single year during the 5-year follow-up period, 72.8 – 99.1% of BC patients had a claim with a 174.x diagnosis code. Another 0 – 11.1% of BC patients did not have a 174.x code but did have a v10.3 code. Of patients without a claim containing 174.x or v10.3, 0 – 0.62% had a claim containing one or more of the diagnosis codes in Table 2 in any given year following treatment.

We conclude that breast cancer patients can be robustly identified within claims databases using diagnosis codes specifically referring to “invasive breast cancer”, including the ICD-9 family of 174.X codes and v10.3, particularly in the treatment and early post-treatment periods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

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 National Cancer Institute; 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.

Funding: This work was funded by grant 1R01HS024936-01 from the Agency for Healthcare Research and Quality

References

1. Riley GF, Administrative and claims records as sources of health care cost data. *Med Care*, 2009 47(7 Suppl 1): p. S51–5. [PubMed: 19536019]
2. Greenfield S, et al., Patterns of care related to age of breast cancer patients. *JAMA*, 1987 257(20): p. 2766–70. [PubMed: 3573271]
3. Mandelblatt JS, et al., Patterns of care in early-stage breast cancer survivors in the first year after cessation of active treatment. *J Clin Oncol*, 2006 24(1): p. 77–84. [PubMed: 16382116]
4. Tyree PT, Lind BK, and Lafferty WE, Challenges of using medical insurance claims data for utilization analysis. *Am J Med Qual*, 2006 21(4): p. 269–75. [PubMed: 16849784]

- Patients with Breast Cancer, but without any utilizations
- ▨ Patients with Breast Cancer with other utilizations
- Patients with Breast Cancer with 233.0, V45.71, 239.3, 238.3, 199.1 Diagnosis Codes, but without 174.x/v10.3 Codes
- ▩ Patients with Breast Cancer with v10.3 Diagnosis Code but without 174.x Codes
- Patients with Breast Cancer with 174.x Diagnosis Codes

Figure 1:

Subgroups of BC patients based on diagnosis codes contained in claims in the initial year of treatment and years 1-5 of follow-up after diagnosis.

Table 1:

Demographics of Patients in the Breast Cancer (BC) and Non-Breast Cancer (non-BC) Groups Without Codes 174.x / v10.3

Total n = 175,454	BC n= 551	nonBC n= 174,903
Age (Year 1) Median (IQR)	80 (74, 86)	73 (68, 80)
Race n (%)		
White	458 (83.1)	145,118 (83.0)
Black	55 (10.0)	14,707 (8.4)
Asian	13 (2.4)	6,849 (3.9)
Hispanic	13 (2.4)	3,715 (2.1)
Other or Unknown	12 (2.2)	4,513 (2.6)
Geographic Region n (%)		
San Francisco	19 (3.5)	5,499 (3.1)
Connecticut	37 (6.7)	9,270 (5.3)
Detroit	30 (5.4)	10,279 (5.9)
Hawaii	11 (2.0)	2,035 (1.2)
Iowa	36 (6.5)	9,702 (5.5)
Seattle	39 (7.1)	7,513 (4.3)
Atlanta	11 (2.0)	4,535 (2.6)
San Jose	14 (2.5)	3,487 (2.0)
Los Angeles	39 (7.1)	11,116 (6.4)
Greater California	100 (18.2)	27,374 (15.7)
Kentucky	42 (7.6)	12,435 (7.1)
Louisiana	26 (4.7)	10,512 (6.0)
New Jersey	74 (13.4)	22,348 (12.8)
Greater Georgia	47 (8.5)	14,655 (8.4)
Other	26 (4.7)	24,143 (13.7)

Table 2.

Diagnoses and codes related to breast cancer or treatment and more commonly associated with breast cancer patients than patients in the general population

Diagnosis Description	ICD-9 Diagnosis Code	O.R.
carcinoma in situ of breast	233.0	41.0
acquired absence of breast and nipple	v45.71	37.0
neoplasm of unspecified nature of breast	239.3	10.0
neoplasm of uncertain behavior of breast	238.3	1.5
malignant neoplasm without specification of site	199.1	1.3