

Validity of Eight Integrated Healthcare Delivery Organizations' Administrative Clinical Data to Capture Breast Cancer Chemotherapy Exposure

Thomas Delate¹, Erin J. Aiello Bowles², Roy Pardee², Robert D. Wellman², Laurel A. Habel³, Marianne Ulcickas Yood⁴, Larissa Nekhlyudov⁵, Katrina A. Goddard⁶, Robert L. Davis⁷, Catherine A. McCarty⁸, Adedayo A. Onitilo⁹, Heather Spencer Feigelson¹⁰, Jared Fremi¹, and Edward Wagner²

Abstract

Background: Cancer Research Network (CRN) sites use administrative data to populate their Virtual Data Warehouse (VDW). However, information on VDW chemotherapy data validity is limited. The purpose of this study was to assess the validity of VDW chemotherapy data.

Methods: This was a retrospective cohort study of women ≥ 18 years with incident, invasive breast cancer diagnosed between January 1999 and December 2007. Pharmacy and procedure chemotherapy data were extracted from each site's VDW. Random samples of 50 patients stratified on trastuzumab, anthracyclines, and no chemotherapy exposure was selected from each site for detailed chart abstraction. Weighted sensitivities and specificities of VDW compared with abstracted data were calculated. Cumulative doses calculated from VDW data were compared with doses obtained from the medical chart review.

Results: The cohort included 13,497 patients with 6,456 (48%) chart review eligible. Patients in the sample ($N = 400$) had a mean age of 65 years. Trastuzumab, anthracycline, and other chemotherapy weighted sensitivities were 95%, 97%, and 100%, respectively; specificities were 99%, 99%, and 93%, respectively; positive predictive values were 96%, 99%, and 55%, respectively; and negative predictive values were 99%, 96%, and 100%. Trastuzumab and anthracyclines VDW mean doses were 873 and 386 mg, respectively, whereas abstracted mean doses were 1,734 and 369 mgs, respectively ($R^2 = 0.14$, $P < 0.01$ and $R^2 = 0.05$, $P = 0.03$, respectively).

Conclusions: Sensitivities and specificities for CRN chemotherapy VDW data were high and dosages were correlated with chart information.

Impact: The findings support the use of CRN data in evaluating chemotherapy exposures and related outcomes. *Cancer Epidemiol Biomarkers Prev*; 21(4); 673–80. ©2012 AACR.

Introduction

Adjuvant chemotherapy is the standard of care in breast cancer when the patient has an increased risk for relapse or progression after initial therapy (1). Randomized clinical trials (RCT) have provided the bulk of chemotherapy

exposure efficacy and safety information. However, the strict inclusion/exclusion criteria of RCTs and standardized treatment protocols may limit the assessments of the consequences of chemotherapy exposure for real-world patients with diverse personal, social, and health care conditions (2). Information on chemotherapy exposures and outcomes may be obtained more cost-effectively with observational studies of well-characterized, large, heterogeneous populations where a range of treatments and subgroups can be studied (3).

Linked Surveillance Epidemiology and End Results (SEER)-Medicare observational data (4) have been validated (5–8) and used in studies of chemotherapy exposures (9–11). While SEER data contain information on cancer type and stage, diagnosis date, some patient characteristics, and whether the patient received first-line therapy (12), special data collection efforts have been undertaken by the National Cancer Institute (Bethesda, MD) to supplement these data with information from random samples of patients with cancer on in-office administrations of specific chemotherapy agents (Patterns of Care studies; ref. 6). Medicare data, which contain

Authors' Affiliations: ¹Pharmacy Department, Kaiser Permanente Colorado, Aurora, Colorado; ²Group Health Cooperative, Group Health Research Institute, Seattle, Washington; ³Kaiser Permanente Northern California, Division of Research, Oakland, California; ⁴Henry Ford Health System, Josephine Ford Cancer Center, Brownstown Twp, Michigan; ⁵Harvard Pilgrim Health Care Institute/Harvard Medical School, Department of Population Medicine and Department of Medicine, Harvard Vanguard Medical Associates, Boston, Massachusetts; ⁶Kaiser Permanente Northwest, Center for Health Research, Portland, Oregon; ⁷Kaiser Permanente Georgia, Center for Health Research Southeast, Atlanta, Georgia; ⁸Essentia Institute of Rural Health, Duluth, Minnesota; ⁹Marshfield Clinic, Marshfield Clinic Research Foundation, Weston, Wisconsin; and ¹⁰Kaiser Permanente Colorado, Institute for Health Research, Denver, Colorado

Corresponding Author: Thomas Delate, Pharmacy Department, Kaiser Permanente Colorado, 16601 E. Centretch Pkwy., Aurora, CO 80011. Phone: 303-739-3538; Fax: 303-739-3574; E-mail: tom.delate@kp.org

doi: 10.1158/1055-9965.EPI-11-1075

©2012 American Association for Cancer Research.

claims for facility and provider billing purposes, are linked to patients' SEER data to provide information on health care services use, including dates and types of chemotherapy administered (12). However, SEER-Medicare data are not without limitations. The linked data typically have limited information on use of oral chemotherapy agents, nondisabled patients <65 years of age, specific chemotherapy dosages, cancer recurrences, and managed care patients (13, 14). In addition, there may be a lag period of 4 years between use of health care services and reporting of data, thus limiting their use for timely investigations (14).

The Cancer Research Network (CRN) is a consortium of 14 nonprofit research centers, based in integrated healthcare delivery organizations, within the HMO Research Network (15). Unlike SEER-Medicare, the CRN's Virtual Data Warehouse (VDW) includes administrative data (i.e., information on members' enrollment, healthcare delivery, and reimbursement for services), managed care patients, and, importantly, information about oral and parenteral chemotherapy including specific agents and dosages (16). Two prior CRN studies have provided some evidence of the validity of the capture of chemotherapy exposure information from the VDW compared with medical record (chart) abstracted data (17, 18). However, because chemotherapy data are collected differently across the CRN sites, it is imperative to assess the validity of the standardized chemotherapy information of CRN. In this study, CRN data were used to conduct a larger and more robust assessment of breast cancer chemotherapy exposure capture. Specifically, the validity of VDW administrative data was assessed in their capture of the receipt of chemotherapy, use of specific chemotherapeutic agents, and dosage information among a cohort of patients with breast cancer who received health care at 8 CRN sites.

Methods

Study design and setting

This was a retrospective, multisite cohort study to evaluate the validity of administrative clinical data to capture chemotherapy exposure. Data were obtained for

patients with breast cancer enrolled at 8 CRN integrated healthcare delivery sites: Group Health Cooperative (GHC), Harvard Pilgrim Health Care and Harvard Vanguard Medical Associates, Henry Ford Health System, Marshfield Clinic, and Kaiser Permanente regions in Colorado, Georgia, Northern California, and the Northwest. These healthcare delivery sites had a combined membership of more than 5 million members in 2008. This study was approved by the GHC Institutional Review Board and 5 other sites that ceded review to GHC and, separately, by the Institutional Review Boards at Marshfield Clinic and Henry Ford Health System.

Patient population

Female patients ages ≥ 18 years, diagnosed with incident invasive (local, regional, or distant summary stages) breast cancer between January 1, 1999 and December 31, 2007, and enrolled in 1 of the 8 CRN healthcare delivery sites at time of diagnosis were included. Patients were enrolled continuously in their respective site during the 12 months before cancer diagnosis (membership gaps of 90 days were permissible; $N = 13,472$). Because of its large patient population, a 10% random sample of eligible women diagnosed from 2001 to 2007 from Kaiser Permanente Northern California were included (chemotherapy data from 1999 and 2000 were incomplete and not included). In addition, Harvard data only included women diagnosed from 2001 to 2006 because of delays in the linkage of their cancer registry data with administrative data.

To maximize the possibility that selected patients were eligible to receive chemotherapy treatment, only women with a tumor size greater than 2.0 cm and/or positive lymph nodes were eligible for chart review ($N = 6,456$). Stratified random samples of 50 patients from each of the 8 sites (total $N = 400$) were selected for detailed medical chart review. If a patient's record was unobtainable, a random substitute patient meeting the same abstraction criteria from the site was reviewed (see Table 1).

Study outcomes

The primary outcomes were weighted sensitivities, specificities, positive predictive values (PPV), and

Table 1. Stratification schema for random sampling of patients for chart reviewing at each of 8 sites

| Exposures | Outcomes | | | | | | Patient count |
|---------------------------------------|-----------------|---------|----------------|---------|----------|---------|---------------|
| | Prevalent CM/HF | | Incident CM/HF | | No CM/HF | | |
| | LVEF | No LVEF | LVEF | No LVEF | LVEF | No LVEF | |
| Both anthracycline and trastuzumab | 1 | 1 | 2 | 2 | 3 | 1 | 10 |
| Either anthracycline or trastuzumab | 1 | 1 | 2 | 2 | 3 | 1 | 10 |
| Neither anthracycline nor trastuzumab | 3 | 3 | 6 | 6 | 9 | 3 | 30 |
| Patient count | 5 | 5 | 10 | 10 | 15 | 5 | 50 |

Abbreviations: CM, cardiomyopathy; HF, heart failure; LVEF, left ventricular ejection fraction measured.

negative predictive values (NPV) of the administrative data for identifying chemotherapy treatment compared with medical record data (the gold standard). Secondary outcomes were a calculation of the prevalences of chemotherapy treatments and an assessment of the sensitivities, specificities, PPVs, and NPVs for patients < and \geq 65 years of age. In addition, an assessment of the variation in the PPVs and NPVs across sites was conducted. Furthermore, an estimation of the cumulative doses of chemotherapy based on administrative data was compared with cumulative doses recorded in patient charts.

Administrative data collection

The CRN uses a federated database (VDW) where each site retains control of their administrative data stored in a common data structure (16). Thus, a programmer at one site can develop programming code that can be run at all sites to extract similar data. The VDW contains patients' health care services procedure data [Healthcare Common Procedure Coding System (HCPCS) codes, including Current Procedural Terminology (CPT) codes, and International Classification of Diseases (ICD)-9 codes] and outpatient pharmacy data, National Drug Codes (NDC). Each CRN site has a data set with all of the NDCs that have ever been used at the site.

The VDW includes tumor registry data with information on each patient's cancer stage at diagnosis, date of cancer diagnosis, age at time of cancer diagnosis, laterality, lymph node involvement, and first-line of cancer treatment. Patient identifiers from the tumor registry were linked with administrative databases to obtain information on pharmacy dispensings, inpatient and outpatient diagnoses and procedures, along with patient characteristics at the time of cancer diagnosis. A comprehensive list of chemotherapy-related HCPCS, CPT, ICD-9, and NDC codes was developed from local, national, and CMS sources. (A list of all codes used in the analyses is available from the authors.) Trastuzumab- and anthracycline-specific treatment data were identified with HCPCS and NDC codes. Nonspecific chemotherapy treatments were identified using CPT, ICD-9, and HCPCS codes (e.g., 96410 - Chemotherapy administration, IV; infusion up to 1 hour). Common programming code was developed, run against each site's VDW (often with local modification because of site-specific differences, such as location and structure of infusion chemotherapy databases), and data were transferred to GHC for analysis.

Procedure and pharmacy administrative data independently were used to calculate cumulative dose for anthracyclines and trastuzumab. Only those patients with evidence of chemotherapy dosing in both their chart and administrative data source were included. Pharmacy dispensing and procedure data were extracted in the 12 months after cancer diagnosis. For each pharmacy dispensing, the days and amount of medication supplied were captured. In addition, information on the count per day of each individual procedure and pharmacy code populated in the administrative data was captured.

Three methods were used to estimate cumulative dose from administrative data. (i) For trastuzumab, a standard dosing of 4 mg/kg loading dose followed by 2 mg/kg follow-up doses was assumed (19). Patient's weight most proximal and before start of chemotherapy was captured and applied to this formula for each specific day (i.e., loading or follow-up) of treatment. These values were then summed across the follow-up period. (ii) For anthracyclines, the concentration of drug in each NDC was multiplied by the amount of drug infused as recorded in the administrative data to obtain the treatment dose [e.g., (NDC 63323-0101-61 = doxorubicin 2 mg/mL) \times (Amount dispensed = 55 mL) = dose of 110 mg]. These values were then summed across the follow-up period. (iii) For trastuzumab and anthracyclines, we summed the number of unique administrations in procedure and pharmacy data for each drug to be used as a proxy for cumulative dose.

Medical record data collection

Each CRN site had access to their patients' electronic and/or paper medical charts. Fifty patients from each site's cohort were selected randomly within stratification groups for detailed chart abstraction (Table 1). Stratification was based on trastuzumab, anthracyclines, and no chemotherapy exposure to ensure representation of women who did and did not receive these drugs. As this study was one aim of a larger study to assess cardiotoxicity of breast cancer chemotherapy treatment, stratification on prevalent (occurring up to 12 months before breast cancer diagnosis), incident (occurring anytime after breast cancer diagnosis through study end), and no heart failure/cardiomyopathy outcomes (based on ICD-9 codes: 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 422.90, 425.4, 425.9, and 428.xx; ref. 20) was undertaken independently to ensure selecting patients who did and did not have one of these diagnoses. Following diagnosis, patient chart data were censored at the time of death, disenrollment from care delivery site, or 1 year after cancer diagnosis, whichever came first.

All abstractors were trained in use of the abstraction tool and blinded to patient sampling stratum. Abstractors reviewed charts depending on availability. Information on patient diagnosis date, stage, characteristics (e.g., age at diagnosis, race/ethnicity), HER2 testing, and chemotherapy treatment (e.g., date of initiation, chemotherapy agent (s), and cumulative dose) were abstracted.

Data analysis

The patient and tumor characteristics were compared among the chart reviewed patients (N = 400; chart reviewed patients), chart-review eligible patients (N = 6,456; chart review eligible patients), and the entire cohort of patients (N = 13,472; entire cohort patients). Because some sites captured chemotherapy in pharmacy databases, procedure data, or both, both pharmacy and procedure administrative data were combined and counted as a single exposure if codes indicated that the same type

of chemotherapy was administered on the same day. We categorized receipt of chemotherapy as ever/never for trastuzumab, any anthracycline (only epirubicin and doxorubicin were dispensed among cohort patients with the vast majority receiving doxorubicin), and other chemotherapy agents.

We calculated prevalences [with 95% confidence intervals (CI)] of ever receiving chemotherapy treatment. Prevalence was calculated as the percentage of patients with chemotherapy treatment information in the data among patients in the respective patient groups. Prevalence was stratified on other chemotherapy, trastuzumab chemotherapy, and anthracycline chemotherapy.

To generalize the chart reviewed patients back to the chart review eligible patients, the inverse probability of verification given the sampling stratification as indicated by the VDW was calculated (Table 1). Calculated weights were scaled to the random sample size ($N = 400$) to provide standard errors relative to the size of the validation cohort. Weighting was only applicable to women eligible for the chart review ($N = 6,456$). Sensitivity was calculated as the percentage of cases with chemotherapy treatment information in both the administrative and medical record data among all patients with chemotherapy treatment information in the medical record data. Specificity was calculated as the percentage of cases with no chemotherapy treatment information in both administrative and medical record data among all patients with no chemotherapy treatment information in the medical record data. Sensitivity, specificity, NPV, and PPV were calculated using weighted logistic regression where the population weights were standardized to reflect the size of the sample ($N = 400$). Data were analyzed using SAS version 9.2 (SAS Institute, Inc.). All statistical tests were 2-sided and P values <0.05 were considered statistically significant.

Sensitivity and specificity were stratified on age (<65 vs. ≥ 65 years) to assess data validity of likely Medicare-eligible and noneligible age categories. The cumulative doses estimated from administrative data for trastuzumab and the anthracyclines were assessed for correlation with doses obtained from chart review using Spearman's correlation coefficient. The correlations were assessed for only those observations where both the chart dose and the administrative dose were not missing.

Results

Among the participating sites, a total of 13,472 patients diagnosed with incident breast cancer during the study period were identified and 6,456 (48%) were eligible for chart review. Of these, a total of 400 (6%) patients were chosen by stratified sampling for manual chart review. Approximately 50% of Chart Reviewed patients were <65 years of age, even though the sample was enhanced for patients diagnosed with heart failure/cardiomyopathy. Most patients were white, recipients of cancer surgery, stage II or III, and had positive lymph node involvement (Table 2).

Trastuzumab, anthracycline, and other chemotherapy exposure was identified in administrative data for 20% ($n = 80$), 38% ($n = 152$), and 21% ($n = 85$), respectively, of the chart reviewed patients. Correspondingly, trastuzumab, anthracycline, and other chemotherapy exposure was noted in the medical charts for 18% ($n = 72$), 42% ($n = 158$), and 10% ($n = 38$), respectively, of these patients. When the administrative data were weighted to the chart review eligible patients ($N = 6,456$), chart reviewed and chart review eligible patient samples had similar exposure prevalences (7% vs. 7%, 55% vs. 51%, and 14% vs. 19% for trastuzumab, anthracycline, and other chemotherapy exposure in the chart reviewed and chart review eligible samples, respectively).

Overall, the weighted sensitivities of administrative data to capture chemotherapy exposure were high ($>92\%$; Table 3). Specificities, PPVs, and NPVs were consistently high ($>93\%$) across data chemotherapy types except for the PPV of other chemotherapy exposure (55%). When examining the sensitivities and specificities of administrative data to capture chemotherapy exposure by likely Medicare-eligible and noneligible age category groups, sensitivities and specificities similarly were high ($>91\%$; Table 4). Across sites, PPVs ranged from 82% to 100%, 98% to 100%, 82% to 100%, and 38% to 94% for trastuzumab only, anthracyclines only, both anthracyclines and trastuzumab, and other chemotherapies, respectively. Across sites, NPVs ranged from 98% to 100%, 92% to 100%, 98% to 100%, and 100% to 100% for trastuzumab only, anthracyclines only, both anthracyclines and trastuzumab, and other chemotherapies, respectively.

Weighted estimates of the cumulative doses varied by chemotherapy type (Table 5). There were a substantial number of patients with limited trastuzumab exposure noted in the chart and, thus, the dose distribution was skewed. Cumulative dose estimates obtained from pharmacy administrative data were modestly correlated and significantly associated with those obtained from medical chart review. Estimates obtained from procedural count administrative data exhibited statistically significant correlations with chart obtained data that appear to capture more accurately cumulative doses.

Discussion

We examined the validity of chemotherapy-related administrative data in capturing information on chemotherapy exposure in a random sample of 400 patients with breast cancer who had received care at 1 of 8 CRN sites across the United States. We found that CRN administrative data were able to identify chemotherapy exposure with a high degree of certainty, including in patients who are not Medicare-eligible. Our findings suggest that the data from the CRN are sufficiently accurate to undertake observational studies of chemotherapy effects in patients with cancer.

Table 2. Characteristics by chart reviewed, chart review eligible, and the entire cohort of patients with breast cancer

| Characteristic | Chart reviewed (N = 400) | Chart review eligible (N = 6,456) | Entire cohort (N = 13,472) |
|--|-----------------------------|--------------------------------------|-------------------------------|
| Mean age at diagnosis (SD, median), y | 65 (14, 65) | 60 (14, 58) | 61 (13, 61) |
| Age ≥ 65 y, n (%) | 207 (52) | 2,326 (36) | 5,521 (41) |
| Year of cancer diagnosis, n (%) | | | |
| 1999–2001 | 119 (30) | 1,903 (30) | 4,144 (31) |
| 2002–2004 | 124 (31) | 2,143 (33) | 4,522 (34) |
| 2005–2006 | 157 (39) | 2,360 (37) | 4,806 (36) |
| Race, n (%) | | | |
| Asian | 13 (3) | 228 (4) | 489 (4) |
| American-Indian/Alaskan native | 1 (<1) | 20 (<1) | 37 (<1) |
| Black/African-American | 67 (17) | 735 (11) | 1,336 (10) |
| Hawaiian/Pacific Islander | 1 (<1) | 11 (<1) | 24 (<1) |
| White | 299 (75) | 5,328 (83) | 11,312 (84) |
| Unknown | 19 (5) | 134 (2) | 274 (2) |
| Worst stage at cancer diagnosis, n (%) | | | |
| I | 48 (12) | 520 (8) | 6,354 (47) |
| II | 182 (46) | 3,683 (57) | 3,711 (28) |
| III | 149 (38) | 1,037 (16) | 1,098 (8) |
| IV | 21 (5) | 244 (4) | 369 (3) |
| Unknown | 0 (0) | 972 (15) | 1,940 (14) |
| Lymph nodes, n (%) | | | |
| Negative | 165 (41) | 2,622 (41) | 9,663 (72) |
| Positive | 235 (59) | 3,834 (59) | 3,834 (28) |
| Surgical treatment, n (%) | | | |
| None | 74 (19) | 333 (5) | 616 (5) |
| Breast conservation | 147 (37) | 2,972 (46) | 7,872 (58) |
| Total mastectomy | 67 (17) | 970 (15) | 1,871 (14) |
| Modified radical mastectomy | 109 (27) | 2,126 (33) | 3,037 (23) |
| Other | 1 (<1) | 17 (<1) | 32 (<1) |
| Missing | 2 (<1) | 38 (<1) | 44 (<1) |

Our findings support previous work by Aiello Bowles and colleagues who reported that the sensitivity of CRN sites' administrative data to identify chemotherapy in patients with ovarian cancer was approximately 90% (17). In addition, our findings are comparable with those reported by Du and colleagues (21) and Warren and colleagues (6) who reported sensitivities of 91% and 88%, respectively, when comparing SEER-Medicare

administrative claims data to medical chart-abstracted data. Our findings are important as they show the validity of CRN administrative data to capture chemotherapy use and provide an alternative/supplement to such data available from SEER-Medicare. We found also a high sensitivity for individual drugs, potentially providing more detailed treatment data than SEER-Medicare because Du and colleagues, using SEER-Medicare data,

Table 3. Weighted estimates of the accuracy of administrative chemotherapy exposure data compared with medical chart abstracted chemotherapy exposure data (N = 400)

| Chemotherapy type | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) |
|-------------------------------|------------------------------|----------------------|-------------------|-------------------------------|
| Trastuzumab only | 94.6% (76.9–98.9) | 99.7% (98.1–99.9) | 96.2% (77.8–99.4) | 99.6% (98.0–99.9) |
| Anthracycline only | 96.7% (93.3–98.4) | 99.5% (96.0–99.9) | 99.6% (96.8–99.9) | 95.8% (91.7–98.0) |
| Anthracycline and trastuzumab | 92.4% (69.4–98.5) | 99.7% (98.1–99.9) | 94.2% (70.3–99.1) | 99.6% (98.1–99.9) |
| Other chemotherapy | 100% (88.1–100) ^a | 93.4% (90.4–95.6) | 55.0% (41.6–67.7) | 100% (98.9–99.4) ^a |

^aExact binomial CIs were computed in cases where accuracy measures were 0% or 100%.

Table 4. Weighted estimates of the sensitivity and specificity of administrative chemotherapy exposure data compared with medical chart abstracted chemotherapy exposure data by age categories (N = 400)

| Chemotherapy type | Sensitivity (95% CI) | | Specificity (95% CI) | |
|-------------------------------|------------------------------|------------------------------|----------------------|-------------------|
| | Age < 65 y | Age ≥ 65 y | Age < 65 y | Age ≥ 65 y |
| Trastuzumab only | 93.6% (73.5–98.7) | 100% (39.8–100) ^a | 99.6% (97.0–99.9) | 99.9% (63.7–100) |
| Anthracycline only | 97.0% (93.3–98.6) | 94.7% (78.1–98.9) | 99.2% (87.8–99.9) | 99.6% (92.7–99.9) |
| Anthracycline and trastuzumab | 91.6% (66.8–98.3) | 100% (15.8–100) ^a | 99.6% (97.1–99.9) | 99.9% (64.1–100) |
| Other chemotherapy | 100% (85.2–100) ^a | 100% (54.1–100) ^a | 94.3% (90.5–96.7) | 91.8% (85.9–95.4) |

^aExact binomial CIs were computed in cases where accuracy measures were 0% or 100%.

confirmed receipt of a specific chemotherapy agent in only 22% of cohort who had received the agent according to chart-abstracted data (21).

We report that among those patients whose chart-abstracted data provided no evidence of receipt of chemotherapy, the specificity of administrative data was more than 99% for trastuzumab and anthracycline exposures. While the specificity for other chemotherapy agents was slightly lower, it still was very high. Our findings are similar to Du and colleagues who reported that SEER-Medicare data correctly identified 99% of patients where chart abstraction data found no evidence of chemotherapy receipt (21) and indicate that CRN administrative data are able to identify with a high degree of certainty patients who did not receive chemotherapy.

With subanalysis by likely Medicare-eligible or not categories (i.e., age <65 vs. ≥65 years), we found very high sensitivities (>93%) and specificities (>91%) in both age groups. The lowest specificity and PPV that we identified were for "other" chemotherapy use among patients ages ≥65 years and among all age groups. Examination of variability across sites revealed small variation in PPVs and NPVs except for other chemotherapies where we found wider variation. Because the goal of our study was to identify the use of anthracyclines and trastuzumab, we did not specify a complete set of possible codes to identify

the "other" agents that may be used in clinical practice. Further enhancement of codes used for the assessment of other chemotherapies is being undertaken by the CRN to provide the most thorough assessment of chemotherapy use.

As future comparative effectiveness, costing, and other retrospective research will need to use cumulative chemotherapy dosing information in their analyses, we attempted to calculate cumulative chemotherapy doses for our study patients from administrative data using methodologies specific to trastuzumab and the anthracyclines. While we found minor correlations between pharmacy administrative data estimates and chart-tabulated cumulative doses, scatter plot analyses (data not shown) indicated a wide distribution of the cumulative doses. However, when using procedural count data to estimate cumulative dose, we found more reasonable correlations between administrative estimated and chart-tabulated data.

The mean doses we do report are within reason for what a patient with typical invasive breast cancer (at 1.8 m²) might receive during a course of treatment for the anthracyclines (e.g., ~a 432-mg cumulative dose over 5 administrations). However, the mean doses we report for trastuzumab appear to be lower than those anticipated for a cumulative dose of trastuzumab (~a 6,000-mg cumulative dose over 50 administrations; ref. 19). The lower

Table 5. Estimates of the cumulative mean chemotherapy dose comparing medical chart abstracted exposure data to administrative chemotherapy exposure data

| Chemotherapy type | Medical chart obtained | Administrative data obtained | Spearman correlation (R^2) | P |
|--------------------------|--|------------------------------|--------------------------------|--------|
| | Mean milligrams (±SD, median) dispensed dose ^a | | | |
| Trastuzumab (N = 74) | 1,734 (±2,252, 0) | 873 (±1,223, 196) | 0.38 (0.14) | <0.001 |
| Anthracyclines (N = 86) | 369 (±178, 416) | 386 (±277, 420) | 0.23 (0.05) | 0.03 |
| | Mean (±SD, median) count of administered cycles ^b | | | |
| Trastuzumab (N = 59) | 15.4 (±10.5, 13) | 20.2 (±14.5, 18) | 0.85 (0.77) | <0.001 |
| Anthracyclines (N = 123) | 4.2 (±1.4, 4) | 5.8 (±3.6, 4) | 0.45 (0.20) | <0.001 |

^aOnly patients with chemotherapy dosing information available in both administrative pharmacy data and chart review data are included.

^bOnly patients with chemotherapy dosing information available in both administrative procedure data and chart review data are included.

cumulative doses we report may be related to the limited follow-up time allowed for by the study design and/or limited patient tolerability compared with those seen in RCTs.

Relatively little, if any, findings have been reported on cumulative dose calculated from SEER-Medicare data. Lamont and colleagues reported the percentage of patients with prespecified billing doses of 5-fluorouracil chemotherapy and the total count of claims for 5-fluorouracil but did not calculate a cumulative doses, nor assess the validity of such doses (5). Similarly, Du and colleagues reported the number of chemotherapy-related claims for patients with breast cancer but not the cumulative chemotherapy dose (21).

Claims data alone might not allow for the calculation of cumulative dose as HCPCS codes for specific chemotherapy agents provide a set milligram amount (e.g., HCPCS J9355: trastuzumab, 10 mg) and, thus, would require multiple claims be present in the data (e.g., for a weekly dose of 110 mg, 11 claims with code J9355 per week would be required to accurately calculate the dose). In addition, HCPCS codes may not provide the fine detail to allow for graduated doses (e.g., 5 mg of trastuzumab). Pharmacy data may provide more detailed cumulative dose estimations with the use of quantity-dispensed information. However, this information was typically missing, incomplete, or inadequate in our pooled data; thus, we relied on algorithms to determine dose. A number of CRN sites are now using the oncology-specific Beacon software component within the HealthConnect (Epic Systems Corp.) electronic medical record. Extracted data from this software provide detailed information on the actual per-patient dose of chemotherapy infused/ingested during an inpatient/in-office cancer care visit. Future evaluations will need to be undertaken to ensure the validity of these data.

Our study had several limitations. Because of resource constraints, we could only abstract a 12-month follow-up period for the chart review data whereas we had a 2-year follow-up for administrative data. Thus, the chart review follow-up may not have been adequate to capture the entire course of chemotherapy for all patients. In addition, some of our patients may have received oncology care outside the healthcare delivery site or the site was a secondary insurance provider; although, this is unlikely as patients in our integrated healthcare delivery systems typically obtain their care at the respective sites. These and the lack of a definitive data on administered dose may have restricted our ability to calculate cumulative dose. However, we used a thorough approach to chart reviewing, the 400 patients were weighted to represent more than 6,000 eligible patients with breast cancer, and we had a

mix of patients who did and did not receive chemotherapy allowing us to calculate sensitivity, specificity, NPV, and PPV.

In conclusion, we report high-quality sensitivity, specificity, PPV, and NPV of CRN administrative data to identify patients with breast cancer with 2+ cm tumors and/or positive lymph nodes who had and had not received chemotherapy, respectively. The sensitivity transcended likely Medicare-eligible age groups. We had mixed results calculating cumulative chemotherapy doses and identifying patients who received other chemotherapies. However, the information obtained provides a context for additional studies to further refine dosing algorithms and identify additional dose and chemotherapy data sources. In total, these findings support the use of CRN administrative data to conduct large-scale population-based studies of both Medicare-eligible and noneligible patients to examine comparative effectiveness of chemotherapy treatment.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: T. Delate, E.J. Aiello Bowles, L.A. Habel, M.U. Yood, L. Nekhlyudov, K.A. Goddard, C.A. McCarty, A.A. Onitilo, J. Freml, E. Wagner

Development of methodology: T. Delate, E.J. Aiello Bowles, R. Pardee, L.A. Habel, M.U. Yood, L. Nekhlyudov, R.L. Davis, C.A. McCarty, A.A. Onitilo, H.S. Feigelson, E. Wagner

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): T. Delate, E.J. Aiello Bowles, L.A. Habel, M.U. Yood, L. Nekhlyudov, K.A. Goddard, R.L. Davis, C.A. McCarty, A.A. Onitilo, H.S. Feigelson, J. Freml

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): T. Delate, E.J. Aiello Bowles, R. Pardee, L.A. Habel, M.U. Yood, L. Nekhlyudov, K.A. Goddard, J. Freml, E. Wagner

Writing, review, and/or revision of the manuscript: T. Delate, E.J. Aiello Bowles, L.A. Habel, M.U. Yood, L. Nekhlyudov, K.A. Goddard, R.L. Davis, C.A. McCarty, A.A. Onitilo, H.S. Feigelson, J. Freml, E. Wagner

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): R. Pardee, L.A. Habel, M.U. Yood, L. Nekhlyudov, R.L. Davis, C.A. McCarty, A.A. Onitilo, J. Freml, E. Wagner

Study supervision: C.A. McCarty, A.A. Onitilo, H.S. Feigelson, E. Wagner

Acknowledgments

The authors thank Priscilla Velentgas, PhD, from Harvard Pilgrim Health Care Institute for her work on designing this study.

Grant Support

This work was supported by a grant from the National Cancer Institute at the NIH (5U19 CA07689-10 to E. Wagner, principal investigator).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received November 14, 2011; revised February 6, 2012; accepted February 7, 2012; published OnlineFirst February 15, 2012.

References

1. Early Breast Cancer Trialists' Collaborative Group (EBCTCG): Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* 2005;365:1687-717.
2. Avorn J. In defense of pharmacoepidemiology—embracing the yin and yang of drug research. *N Engl J Med* 2007;357:2219-21.
3. Spigel DR. The value of observational cohort studies for cancer drugs. *Biotechnol Healthc* 2010;7:18-24.

4. National Cancer Institute. SEER-Medicare Linked Database. [cited 2011 Jan 20]. Available from: <http://healthservices.cancer.gov/seermedicare>.
5. Lamont EB, Lauderdale DS, Schilsky RL, Christakis NA. Construct validity of Medicare chemotherapy claims: the case of 5FU. *Med Care* 2002;40:201–11.
6. Warren JL, Harlan LC, Fahey A, Virnig BA, Freeman JL, Klabunde CN, et al. Utility of the SEER-Medicare data to identify chemotherapy use. *Med Care* 2002;40 Suppl:IV-55–61.
7. Potosky AL, Riley GF, Lubitz JD, Mentnech RM, Kessler LG. Potential for cancer related health services research using a linked Medicare-tumor registry database. *Med Care* 1993;31:732–48.
8. McClish DK, Penberthy L, Whittemore M, Newschaffer C, Woolard D, Desch CE, et al. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol* 1997;145:227–33.
9. Dobie SA, Baldwin LM, Dominitz JA, Matthews B, Billingsley K, Barlow W. Completion of therapy by Medicare patients with stage III colon cancer. *J Natl Cancer Inst* 2006;98:610–9.
10. Du XL, Chan W, Giordano S, Geraci JM, Delclos GL, Burau K, et al. Variation in modes of chemotherapy administration for breast carcinoma and association with hospitalization for chemotherapy-related toxicity. *Cancer* 2005;104:913–24.
11. Elkin EB, Hurria A, Mitra N, Schrag D, Panageas KS. Adjuvant chemotherapy and survival in older women with hormone receptor-negative breast cancer: assessing outcome in a population-based, observational cohort. *J Clin Oncol* 2006;24:2757–64.
12. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care* 2002;40 Suppl:IV-3–18.
13. Bach PB, Guadagnoli E, Schrag D, Schussler N, Warren JL. Patient demographic and socioeconomic characteristics in the SEER-Medicare database: applications and limitations. *Med Care*. 2002;40 Suppl:IV-19–25.
14. Potosky AL. The Linked SEER-Medicare Data and Cancer Effectiveness Research. [cited 2011 Jan 20]. Available from: <http://www.iom.edu/~media/Files/Activity%20Files/Disease/NCPF/2009-OCT-5/Potosky-TheLinkedSEER-MedicareDataandCancerEffectivenessResearch.pdf>.
15. Wagner EH, Greene SM, Hart G, Field TS, Fletcher S, Geiger AM, et al. Building a research consortium of large health systems: the Cancer Research Network. *J Nat Cancer Inst Monogr* 2005;35:3–11.
16. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monograph* 2005;35:12–25. Available from: <http://crn.cancer.gov/>.
17. Aiello Bowles EJ, Tuzzio L, Ritzwoller DP, Williams AE, Ross T, Wagner EH, et al. Accuracy and complexities of using automated clinical data for capturing chemotherapy administrations: implications for future research. *Med Care* 2009;47:1091–7.
18. Buist DSM, Chubak J, Prout M, Yood MU, Bosco JLF, Thwin SS, et al. Referral, receipt and completion of chemotherapy in early stage breast cancer patients aged 65 and older at high-risk of breast cancer recurrence. *J Clin Oncol* 2009;27:4508–14.
19. NCCN Guidelines™ Version 2.2011 Invasive Breast Cancer: National Comprehensive Cancer Network. [cited 2011 Jan 21]. Available from: http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf.
20. Go AS, Lee WY, Yang J, Lo JC, Gurwitz JH. Statin therapy and risks for death and hospitalization in chronic heart failure. *JAMA* 2006;296:2105–11.
21. Du XL, Key CR, Dickie L, Darling R, Geraci JM, Zhang D. External validation of Medicare claims for breast cancer chemotherapy compared with medical chart reviews. *Med Care* 2006;44:124–31.