



Published in final edited form as:

Med Care. 2009 October ; 47(10): 1091–1097. doi:10.1097/MLR.0b013e3181a7e569.

Accuracy and complexities of using automated clinical data for capturing chemotherapy administrations: implications for future research

Erin J. Aiello Bowles, MPH¹, Leah Tuzzio, MPH¹, Debra P. Ritzwoller, PhD², Andrew Williams, PhD³, Tyler Ross, MA¹, Edward H. Wagner, MD, MPH¹, Christine Neslund-Dudas, MPH⁴, Andrea Altschuler, PhD⁵, Virginia Quinn, PhD⁶, Mark Hornbrook, PhD⁷, and Larissa Nekhlyudov, MD, MPH⁸

¹Group Health Center for Health Studies, Seattle, WA

²Institute for Health Research, Kaiser Permanente Colorado, Denver, CO

³Kaiser Permanente Hawaii, Honolulu, HI

⁴Henry Ford Health System, Detroit, MI

⁵Kaiser Permanente Northern California, Oakland, CA

⁶Kaiser Permanente Southern California, Pasadena, CA

⁷Kaiser Permanente Northwest, Portland, OR

⁸Department of Ambulatory Care and Prevention, Harvard Medical School/Harvard Pilgrim Health Care, Boston, MA

Abstract

Background—Chemotherapy data are important to almost any study on cancer prognosis and outcomes. However, chemotherapy data obtained from tumor registries may be **incomplete**, and abstracting chemotherapy directly from medical records can be expensive and time consuming.

Methods—We evaluated the accuracy of using automated clinical data to capture chemotherapy administrations in a cohort of 757 ovarian cancer patients enrolled in seven health plans in the HMO Cancer Research Network. We calculated sensitivity and specificity with 95% confidence intervals (CI) of chemotherapy administrations extracted from three automated clinical data sources (Health Care Procedure Coding System [HCPCS], National Drug Codes, and International Classification of Diseases) compared to tumor registry data and medical chart data.

Corresponding author and reprint requests: Erin J. Aiello Bowles, MPH Group Health Center for Health Studies 1730 Minor Ave, Suite 1600 Seattle WA, 98101 Ph: 206–287–2708 Fax: 206–287–2485 Email: aiello.e@ghc.org. Leah Tuzzio, MPH, 1730 Minor Ave, Suite 1600, Seattle, WA 98101. Ph: 206–287–2109; fax: 206–287–2871; email: Tuzzio.l@ghc.org. Debra P. Ritzwoller, PhD, 1065 E. Harvard Ave, Suite 300, Denver, CO, 80231. Ph: 303–614–1317; fax: 303–614–1305; email: debra.ritzwoller@kp.org. Andrew Williams, PhD, 501 Alaska St, Suite 201, Honolulu, HI, 96817. Ph: 808–432–4756; fax 808–432–4785; email: Andrew.e.williams@kp.org. Tyler Ross, MA, 1730 Minor Ave, Suite 1600, Seattle, WA 98101. Ph: 206–287–2927; fax: 206–287–2871; email: ross.t@ghc.org. Edward H. Wagner, MD, MPH, 1730 Minor Ave, Suite 1600, Seattle, WA 98101. Ph: 206–287–2877; fax: 206–287–2138; email: wagner.e@ghc.org. Christine Neslund-Dudas, MPH, 1 Ford Place, Detroit, MI 48202. Ph: 313–874–5472; fax: 313–874–6656; email: cdudas1@hfhs.org. Andrea Altschuler, PhD, 2000 Broadway, Oakland, CA 94612. Ph: 510–891–3560; fax: 510–891–3606; email: andrea.altschuler@kp.org. Virginia Quinn, PhD, 100 South Los Robles, Pasadena, CA 91188. Ph: 626–564–3467; fax: 626–564–3430; email: Virginia.p.quinn@kp.org. Mark Hornbrook, PhD, 3800 North Interstate Ave, Portland, OR 97227. Ph: 503–334–6746; fax: 503–334–2428; email: mark.c.hornbrook@kpchr.org. Larissa Nekhlyudov, MD, MPH, 133 Brookline Ave, 6th Floor, Boston, MA 02215. ph: 617–509–9983; fax: 617–8598112; email: larissa_nekhlyudov@harvardpilgrim.org.

Results—Sensitivity of all three data sources varied across health plans from 79.4–95.2% when compared with tumor registries, and 75.0–100.0% when compared with medical charts. The sensitivities using a combination of three data sources were 88.6% (95%CI 85.7–91.1) compared with tumor registries and 89.5% (78.5–96.0) compared with medical records; specificities were 91.5% (86.4–95.2) and 90.0% (55.5–99.7), respectively. There was no difference in accuracy between women aged <65 and ≥65 years. Using one set of codes alone (e.g., HCPCS alone) was insufficient for capturing chemotherapy data at most health plans.

Conclusions—While automated data systems are not without limitations, clinical codes used in combination are useful in capturing chemotherapy more comprehensively than tumor registry and without the need for costly medical record abstraction.

Keywords

validation of automated clinical data; chemotherapy; medical chart; tumor registry; ovarian cancer

Introduction

The number of cancer survivors in the U.S. has increased steadily over the past 30 years.(1) This growing population has stimulated new research questions about the long-term effects of chemotherapy on cancer recurrence, mortality, and health-related quality of life.(2) Observational studies on chemotherapy outcomes are critical to help inform patients and physicians about the long-term risks and benefits associated with treatment.(3) However, population-based research on chemotherapy and outcomes is not common due to the difficulty in obtaining accurate chemotherapy exposure information.(4) Exploring new methods to accurately obtain chemotherapy data could make this kind of research much more feasible.

Most studies that have collected chemotherapy data have relied on one of three data sources: 1) patient medical records; 2) local or national tumor registries; and/or 3) the SEER-Medicare dataset, which links the SEER (Surveillance, Epidemiology, and End Results) tumor registries to Medicare claims files. While medical records contain detailed data on chemotherapy administrations including the route, dose, and type of drug administered, they are expensive and time-consuming to abstract, thereby rendering population-based studies infeasible. SEER and other tumor registries collect the majority of their treatment data from inpatient medical record abstraction. While SEER data are relatively inexpensive and easy to obtain for research purposes, information on the use of adjuvant chemotherapy is often underreported because treatment usually occurs in the outpatient setting.(5-7) To improve efficiency and accuracy of population-based cancer research, it is important to explore other data sources and methods of obtaining chemotherapy data.

Automated clinical data include billing and claims records, which contain diagnosis, drug, and procedure codes assigned at the time of chemotherapy administration. Together, these data provide information on the administration of the drug, the type of drug given, and the associated costs. Several studies have evaluated the accuracy of automated clinical data for capturing chemotherapy.(4,7) Warren et al. evaluated Medicare data with Health Care Procedure Coding System (HCPCS) codes (including Current Procedure Terminology [CPT-4] codes) and International Classification of Diseases (ICD-9-CM) codes for any chemotherapy administrations and found a sensitivity of 88% compared to a gold standard of medical charts.(4) Similarly, Du et al. estimated good sensitivity (91%) and specificity (95%) when comparing automated clinical data from SEER-Medicare files to medical charts for 1,228 breast cancer patients.(8) While using automated data as a method to extract chemotherapy administrations shows promise, to our knowledge, this method has not been evaluated in non-Medicare populations. Automated clinical data from integrated health plans might be one source from

which researchers could study chemotherapy data among a younger population, who may especially benefit from research on long-term chemotherapy effects.

We quantified sensitivity and specificity of chemotherapy administrations from automated clinical data relative to two gold standards: tumor registries and medical charts. Our patient sample included 757 ovarian cancer patients enrolled in one of seven health plans in the HMO Cancer Research Network (CRN). The CRN health plans have automated data on cancer diagnoses and outcomes, co-morbid diagnoses, health plan enrollment, outpatient pharmacy administrations, and healthcare utilization.(9) The inclusion of patients under 65 years of age and the availability of pharmacy data represent at least two advantages of these data over the SEER-Medicare files.(10) Therefore, our study's findings may be applied to future studies within US health plans or other settings with access to standard automated clinical data seeking to efficiently conduct larger, population-based studies on chemotherapy treatment and outcomes.

Methods

We conducted this study within the HMO Cancer Research Network (CRN), a consortium of 14 research organizations affiliated with non-profit integrated healthcare delivery systems and the National Cancer Institute (NCI). The participating health plans in this study were Group Health, Harvard Pilgrim Health Care, Henry Ford Health System, and Kaiser Permanente regions in Colorado, Northwest, Hawaii, Northern California, and Southern California. All health plans had access to their patient data collected by their regional cancer registry, including Surveillance, Epidemiology, and End Results (SEER), state, and/or local tumor registries.(9) Group Health, Henry Ford Health System, and the Kaiser Permanente regions of Colorado, Hawaii, Northern California, and Southern California all report to SEER and respective state registries; Harvard Pilgrim Health Care and Kaiser Permanente Northwest only report to state registries. These registries are all population-based and meet the high quality reporting standards for SEER or state registries.(9,11) Each registry abstracts and combines data from multiple sources so that all diagnostic and treatment information are available in one record, even if part of the treatment occurred at a hospital outside of the HMO. All health plans had complete tumor registry data within one year of diagnosis.

This study was originally conducted to assess the diffusion of intraperitoneal (IP) chemotherapy for ovarian cancer in response to a clinical announcement published by NCI in January 2006.(12) The methods and results below describe the validation study conducted to assess the accuracy of automated chemotherapy data from several sources. Eight health plans were included in the original IP chemotherapy study, but we excluded one health plan from this validation study because it had limited access to the required automated clinical data needed for this report. All health plans obtained approval for this study from their local Institutional Review Boards.

Study Population

We identified all women over 18 years of age with an incident diagnosis of ovarian cancer between January 1, 2004 and June 30, 2006 from the seven health plans (N=1,068). We excluded those cases with ovarian histology other than epithelial, borderline, or other ovarian; non-primary diagnoses; and cases diagnosed with stage 0 disease in order to meet the primary aim of the study (assessing the diffusion of intraperitoneal chemotherapy). The remaining sample included 757 cases across seven sites. To accommodate lags in capturing automated data from health plans, the data pull took place between May 16, 2007 and November 8, 2007.

Data were aggregated using the CRN Virtual Data Warehouse (VDW), established by the CRN for efficient collaboration and pooling of automated data across sites using a standard data

dictionary.(11) The VDW is virtual in that it is a distributed data system where the health plans retain local control of their data, but a programmer at one site can write a program than can be run at all sites.(11) Using the CRN's VDW, we collected automated data on patient characteristics including demographics (age, race, ethnicity, vital status), health plan enrollment, tumor characteristics (stage and histology obtained from population-based tumor registries), and treatment received (surgery, radiation, chemotherapy, and hormone treatment obtained from tumor registries). Utilization files include automated clinical data associated with inpatient and outpatient encounters, types of procedure and diagnostic codes including International Classification of Diseases (ICD-9-CM) codes for procedures and diagnoses, or Healthcare Common Procedure Coding System (HCPCS) codes for procedures. HCPCS codes incorporate the American Medical Association's Common Procedure Terminology codes (CPT-4) for medical, surgical, laboratory, and imaging procedures, as well as additional codes for other procedures reimbursed by Medicare.

Ascertainment of Chemotherapy from Automated Clinical Data

We identified chemotherapy data using three types of codes: HCPCS codes including CPT-4 procedure codes from inpatient and outpatient data; National Drug Codes (NDCs) from pharmacy data; and ICD-9-CM codes (including procedures and diagnoses) from inpatient and outpatient data. We did not include revenue center codes because not all participating sites had access to these data. The HCPCS we included were level I HCPCS (CPT-4) 96400–96549 and level II HCPCS J9000–J9999 and Q0083–Q0085; ICD-9-CM included diagnostic codes 00.10 and 99.25, and procedure code V58.1. These codes primarily identify infusion procedures and cancer chemotherapy. The NDCs identify anti-neoplastic agents and are listed in Appendix 1, <http://links.lww.com/A1256>. We used these codes to identify treatments that occurred in the period between January 1, 2004, and June 30, 2006.

Within each set of codes (HCPCS, NDCs, and ICD-9-CM), we collapsed the codes to create dichotomous variables indicating whether each woman had any chemotherapy or none. We also combined all three types of codes to create an overall dichotomous variable indicating any chemotherapy treatment or none.

Methods for chart abstraction

We selected 81 charts across the seven health plans for chart abstraction. In accordance with the primary aim of the original study, we used a within-site stratified random sampling procedure to enhance the probability of selecting the charts of patients who received IP treatment. First, we selected all charts (N=12) for cases containing a CPT-4 code for IP therapy (96445) or use of an IP catheter (49419 or 49420). Second, we randomly selected the remaining charts (N=69) from cases diagnosed with stage 2B or greater disease (the patient group targeted in the NCI clinical announcement). We stratified chart selection to cover incident ovarian cancer diagnosed during the periods before and after the date of the clinical announcement: half from incident diagnoses between January 2004 and December 2005, and the other half from incident diagnoses between January and June 2006. After review, we excluded charts for 11 women without primary diagnoses or diagnosed with histology other than epithelial or borderline (final chart sample = 70).

Chart abstractors at each site reviewed both electronic and paper medical records, including all inpatient and outpatient visits with multiple providers. The chart abstraction instrument for this study was based on one being developed for the National Comprehensive Cancer Network (NCCN) provided by Dr. Jane Weeks (grant number CA006516–43S3 from the NCI, PI Benz). We abstracted information on the following items: ovarian cancer diagnosis stage and histology, treatment received (including surgery, chemotherapy, radiation, and hormone treatment), and recurrence/survival outcomes. Instructions were provided on the abstraction

form and several abstractor sessions were held by telephone conference. In order to minimize potential bias from the cases that were selected on the basis of a CPT-4 code for IP chemotherapy, chart abstractors were blinded to the chemotherapy status from the automated clinical data at the time of review. For each chart, we abstracted chemotherapy treatment through June 30, 2006.

Statistical Analysis

We described the total population of ovarian cancer cases by demographic and tumor characteristics and compared the distribution of characteristics among the entire cohort to the subset of cases with medical record abstraction data. In order to evaluate the accuracy of chemotherapy administrations from automated clinical data, we coded women as having received chemotherapy if they had a HCPCS, NDC, or ICD-9-CM code indicating one or more administrations. We coded women as having no chemotherapy if they had no administrations in any one of the three automated clinical data sources.

We calculated the sensitivity and specificity of automated clinical data for ascertaining chemotherapy information compared to our two gold standards – tumor registries and medical records. The sensitivity (calculated with 95% confidence intervals [CI]) was the proportion of cases with chemotherapy administrations in both automated clinical data and each gold standard among all cases with chemotherapy according to each gold standard. Specificity (with 95% CI) was the proportion of cases without any chemotherapy administrations in the automated clinical data and each gold standard among all women with no chemotherapy according to each gold standard. We stratified our sensitivity and specificity analyses to evaluate whether these measures varied across sites or by patient characteristics such as age and stage at diagnosis. All analyses were conducted in StataSE 9.0 (College Station, TX).

Results

The characteristics of the patient population from our automated clinical data are outlined in Table 1. The majority of the 757 women with primary, invasive, epithelial ovarian cancer were over 50 years old, white, and long-term enrollees at their health plan. More than half were diagnosed with stage III or IV disease. According to the tumor registry data, 82% of women received surgery and 77% received chemotherapy. The subset of chart-reviewed cases (n=70) was similar to the overall study population with the exception of stage, diagnosis dates, and chemotherapy treatment as these were the covariates we used to select women for chart abstraction.

Table 2 shows the sensitivity and specificity of each type of automated clinical data code used to extract chemotherapy data across seven CRN health plans compared to tumor registry and medical record data. When evaluating each type of code alone, HCPCS codes had a higher sensitivity than NDCs or ICD-9-CM codes alone compared with both tumor registry (64.0%, 95%CI 59.9–67.9) and medical record (66.7%, 95%CI 52.9–78.6) data. Combining all three codes resulted in a higher sensitivity than any one or two types of codes alone (compared to tumor registries: 88.6%, 95%CI 85.7–91.1; compared to medical records: 89.5%, 95%CI 78.5–96.0). However, as sensitivity increased, specificity decreased. Specificity was 91.5% (95% CI 86.4–95.2) for three codes combined when compared to tumor registry data and 90.0% (95%CI 55.5–99.7) when compared to medical record data. Specificity was slightly higher for individual codes than combined codes when compared to tumor registry data.

Figure 1 shows the sensitivity of each type of automated clinical data code compared to tumor registries stratified by each of the seven CRN health plans. While some sites had high sensitivities for capturing chemotherapy using HCPCS codes alone (sites 1, 4, 6, 5, and 3; Figure 1a), others had high sensitivities when using NDC codes alone (sites 2, 5, and 3; Figure

1b). Sensitivity ranged across sites from 79.4–95.2% when comparing all three codes to tumor registries (Figure 1d). We found similar results for sensitivity when comparing automated clinical data to medical records across health plans (Figure 2). When evaluating all three types of codes together, the range of sensitivities across the seven health plans was wider (down to 75.0%) although two health plans (sites 1 & 2) had 100% sensitivity (Figure 2d).

There was no difference in sensitivity between women diagnosed before age 65 and women ≥ 65 (88.7%, 95%CI 85.0–91.8; and 88.5%, 95%CI 83.5–92.4, respectively) when comparing all three codes combined to tumor registries (Figure 3a). Specificity in women <65 was 92.1% (95%CI 83.6–97.0) compared with 91.1% (95%CI 83.8–95.8) in women ≥ 65 (data not shown). Accuracy of automated data did not differ by age when validated against medical charts (Figure 3b). Among women <65 , the sensitivity and specificity were 87.1% (95%CI 70.2–96.4) and 100% (95%CI 47.8–100.0), respectively (specificity data not shown). Among women ≥ 65 , the sensitivity and specificity were 92.3% (95%CI 74.9–99.1) and 80% (95%CI 28.4–99.5), respectively (specificity data not shown). We noted no differences in sensitivity or specificity by stage at diagnosis (data not shown).

Discussion

We found that automated clinical data using HCPCS, NDC, and ICD-9-CM codes in combination had 88.6% sensitivity and 91.5% specificity for capturing chemotherapy administrations when compared with tumor registry data across seven health plans participating in the CRN. When we compared automated clinical data for capturing chemotherapy data with medical charts, the sensitivity was 89.5% and specificity was 90.0%. To our knowledge, this is the first study to evaluate these codes for ascertaining chemotherapy data in a population that includes some non-Medicare patients. We did not find any difference in the accuracy of these codes between women <65 and 65 years and older. However, our results showed that one automated clinical data source alone (e.g. HCPCS alone) may not be sufficient for extracting chemotherapy data for research purposes.

Du et al. and Warren et al. found sensitivities of 91% and 88%, respectively, when comparing automated clinical codes from SEER-Medicare data to medical records.^(4,8) These sensitivities are similar to those noted in our study. However, we initially excluded one health plan from our study that might have lowered our sensitivity and specificity had it been included. These results suggest that these codes are valid for research use when you have reliable access to standard automated clinical codes. As previously noted, further investigation into the outlying health plan revealed limited access to all automated clinical data.

There are two major differences between our study and previous work. First, we included NDCs from pharmacy data, which are not available in SEER-Medicare data. Including NDCs increased the sensitivity of using automated clinical data by 8.9 percentage points when comparing to tumor registries; however, the specificity decreased by 1.7%. There is always a tradeoff between sensitivity and specificity – as one increases, the other usually decreases. Other studies that plan to use automated clinical data to identify chemotherapy need to weigh the importance of maximizing sensitivity or specificity when choosing their data sources. Second, we were able to evaluate the sensitivity and specificity of automated clinical data among cancer patients younger than 65. These data are also not available in SEER-Medicare. Our results showed that the sensitivity and specificity of automated clinical data were almost identical for women <65 and ≥ 65 years of age. This is reassuring for future studies suggesting there do not appear to be limitations in using automated clinical data for the population not covered by Medicare.

When evaluating sensitivity and specificity in our study, it is important to note the following. First, when we used medical records as the gold standard, the sensitivities were very slightly higher than those when using tumor registry data as the gold standard. This is because we selected medical records for abstraction based on prior knowledge of CPT-4 codes for intraperitoneal chemotherapy in the population (as the primary goal of the study was to evaluate diffusion of IP chemotherapy). Although these codes were rare (N=12 cases with a code for IP chemotherapy or catheter placement), this selection criteria may have slightly inflated our sensitivity estimates using medical records as the gold standard.

Second, tumor registries are often thought to have incomplete ascertainment of chemotherapy data, raising questions about their validity as a gold standard.(4,6,7) While Cress et al. reported that tumor registries have more complete chemotherapy data on patients receiving treatment at HMOs compared to fee-for-service hospitals, we may still have some underreporting in our study.(6) It is possible that the automated clinical data may actually be more complete than the tumor registry data. If so, the false positives (i.e. chemotherapy captured by automated clinical data but not in the tumor registry) across the seven CRN health plans (1 – specificity [91.5%] or 8.5%) may actually be true positives. In other words, the tumor registries likely missed capturing the chemotherapy treatment in these cases whereas the automated clinical data did not. Five out of seven health plans had 100% specificity when comparing automated clinical data to medical charts (which should not have missed any chemotherapy administrations), adding support to this explanation.

The false negatives (1- sensitivity) in this study warrant further discussion. Our results suggest that the automated clinical codes missed 11.4% of chemotherapy administrations captured by tumor registries and 10.0% recorded in medical records. While it is unclear why automated clinical data may be missing these treatment records, we offer two potential explanations. Because our primary aim for this study was to examine the use of IP chemotherapy, over one quarter of the women (28.6%) selected for the medical record review were diagnosed between January and June 2006. Since our data were pulled between May and November 2007, it is possible that the sensitivity of automated clinical data for capturing chemotherapy administrations was lower for the chemotherapy data for these women were incomplete. It is also possible that each of the health plans have a small number of clinics and outside referral care sources that use non-standardized codes or disparate data systems for recording chemotherapy data that are not included in the VDW. We did not include revenue center codes in our analysis, which could have reduced the number of false negatives at sites with these data. We could not account for these potentially missing data in our study and this limitation needs to be taken into consideration in any analysis of health plan data.

This study has several strengths and limitations. Using the CRN VDW, we were able to extract automated data on a large cohort of ovarian cancer cases. Given the rarity of this disease, this study would not have been feasible without collaboration across multiple institutions. Using the VDW, we were able to gather and analyze a large amount of data quickly and cost-effectively. In addition, the results of this study are likely generalizable to other studies and settings with access to automated clinical data and may be particularly useful for studying for populations under age 65. They may also be generalizable to other cancers if chemotherapy data is coded with standard HCPCS, NDC, or ICD-9 codes in automated clinical data. This study was limited by the number of medical record reviews. Because this number was small, we were unable to conduct additional analyses including the sensitivity and specificity of automated clinical data for identifying specific chemotherapy agents or the number of cycles.

In summary, this is the first study to our knowledge to evaluate automated clinical data for ascertaining chemotherapy administrations for the population under 65. While we found substantial variation in the use of different types of automated codes across health plans, the

variation was greatly reduced when we used all three codes combined. Automated data systems are not without limitations, and studies have shown that HCPCS, NDC, and ICD-9-CM codes are no exception.⁽¹³⁾ However, when these codes are used in combination it is possible to capture chemotherapy more comprehensively than tumor registries and without the need for costly medical record abstraction. This study has furthered our ability to understand the benefits and limitations of the health plan automated clinical data in capturing chemotherapy information. Validation of automated clinical data across numerous U.S. health plans may support large-scale population-based studies examining the utilization, costs, and outcomes of chemotherapy, particularly in younger cancer patients who may benefit the most from research on long-term chemotherapy effects.

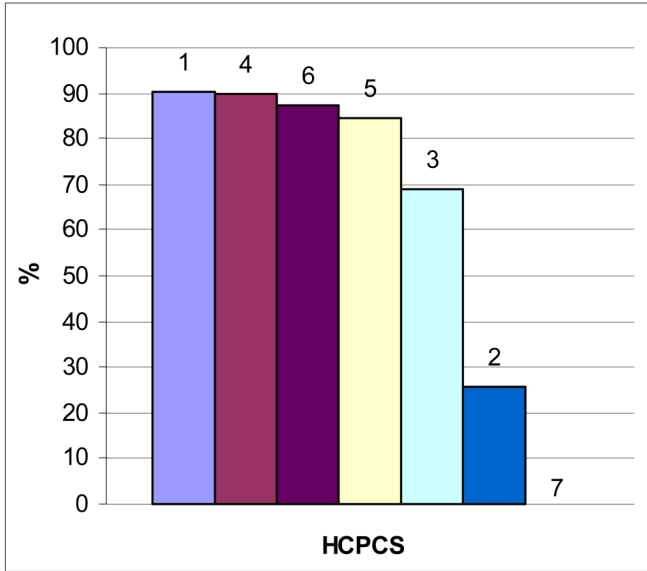
Acknowledgments

Funding: This study was funded by Grant No. U19-CA79689 from the National Cancer Institute for the Cancer Research Network.

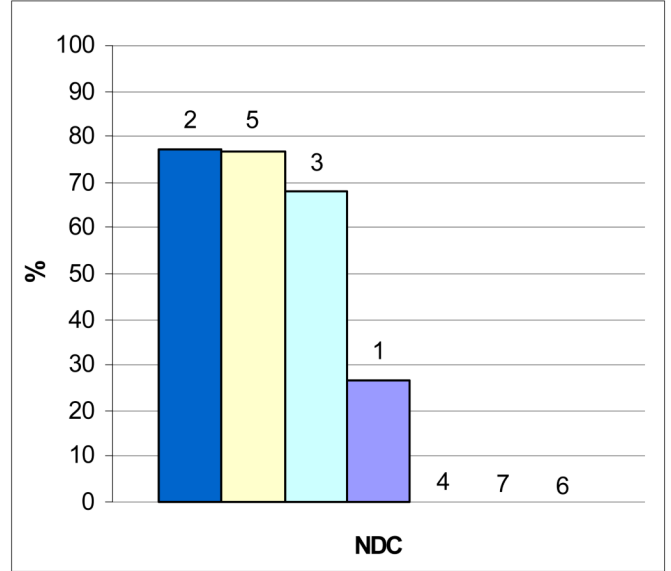
References

1. Ries LAG, MD.; Krapcho, M.; Mariotto, A.; Miller, BA.; Feuer, EJ.; Clegg, L.; Horner, MJ.; Howlader, N.; Eisner, MP.; Reichman, M.; Edwards, BK., editors. SEER Cancer Statistics Review, 1975–2004. National Cancer Institute; Bethesda, MD:
2. NCI Plans and Priorities. Cancer Survivorship. [July 23, 2008]. Available at: <http://plan2004.cancer.gov/public/survivor.htm>.
3. Geiger AM, Buist DS, Greene SM, et al. Survivorship research based in integrated healthcare delivery systems: the Cancer Research Network. *Cancer* 2008;112:2617–2626. [PubMed: 18428194]
4. Warren JL, Harlan LC, Fahey A, et al. Utility of the SEER-Medicare data to identify chemotherapy use. *Med Care* 2002;40:IV, 55–61.
5. Du XL, Key CR, Dickie L, et al. Information on chemotherapy and hormone therapy from tumor registry had moderate agreement with chart reviews. *J Clin Epidemiol* 2006;59:53–60. [PubMed: 16360561]
6. Cress RD, Zaslavsky AM, West DW, et al. Completeness of information on adjuvant therapies for colorectal cancer in population-based cancer registries. *Med Care* 2003;41:1006–1012. [PubMed: 12972840]
7. Roetzheim RG, Chirikos TN, Wells KJ, et al. Managed care and cancer outcomes for Medicare beneficiaries with disabilities. *Am J Manag Care* 2008;14:287–296. [PubMed: 18471033]
8. Du XL, Key CR, Dickie L, et al. External validation of medicare claims for breast cancer chemotherapy compared with medical chart reviews. *Med Care* 2006;44:124–131. [PubMed: 16434911]
9. Wagner EH, Greene SM, Hart G, et al. Building a research consortium of large health systems: the Cancer Research Network. *J Natl Cancer Inst Monogr* 2005:3–11. [PubMed: 16287880]
10. Warren JL, Klabunde CN, Schrag D, et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care* 2002;40:IV, 3–18.
11. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr* 2005:12–25. [PubMed: 16287881]
12. Armstrong DK, Bundy B, Wenzel L, et al. Intraperitoneal cisplatin and paclitaxel in ovarian cancer. *N Engl J Med* 2006;354:34–43. [PubMed: 16394300]
13. Ramsey SD, Martins RG, Blough DK, et al. Second-line and third-line chemotherapy for lung cancer: use and cost. *Am J Manag Care* 2008;14:297–306. [PubMed: 18471034]

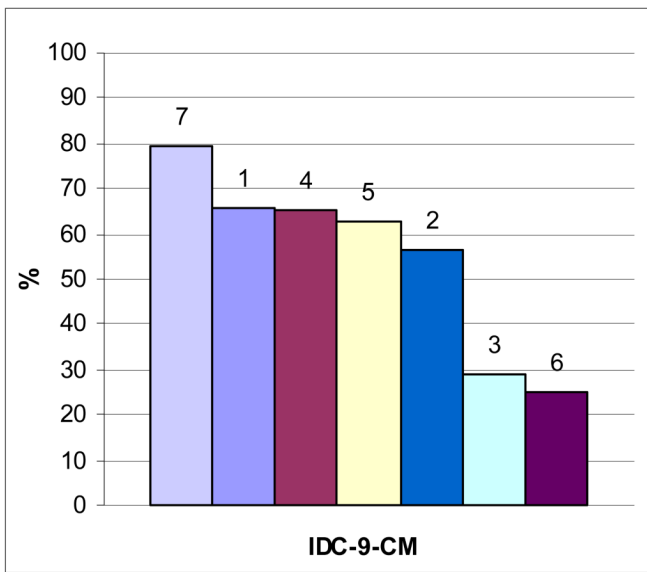
1a. Sensitivity of HCPCS codes



1b. Sensitivity of NDCs



1c. Sensitivity of ICD-9-CM codes



1d. Sensitivity of HCPCS + NDC + ICD-9-CM

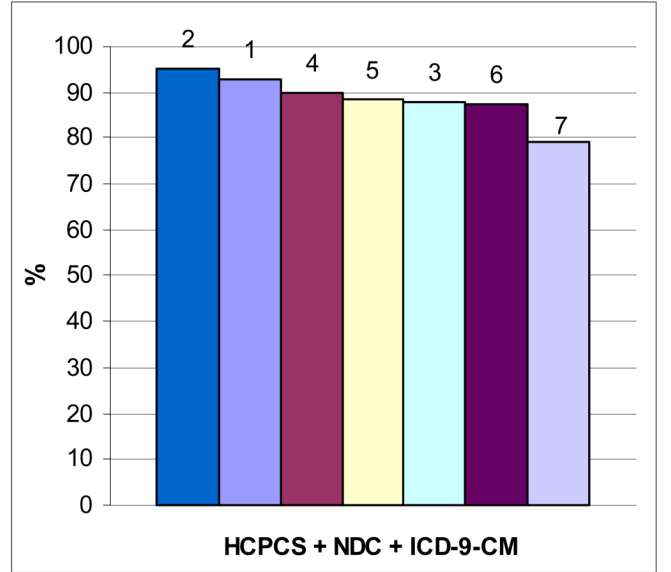
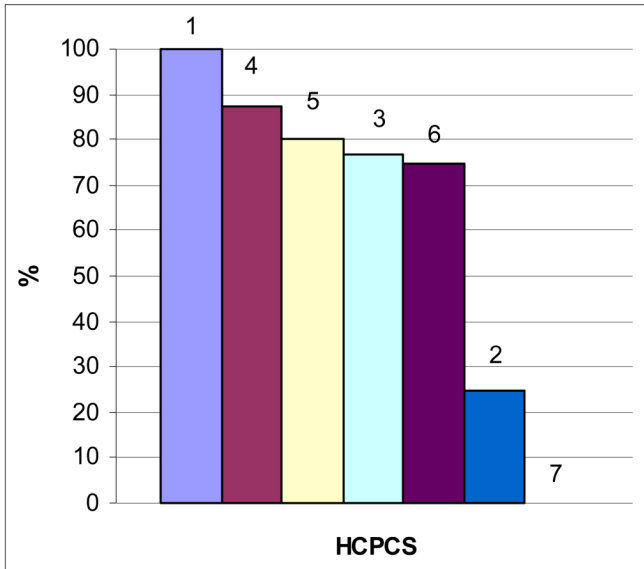


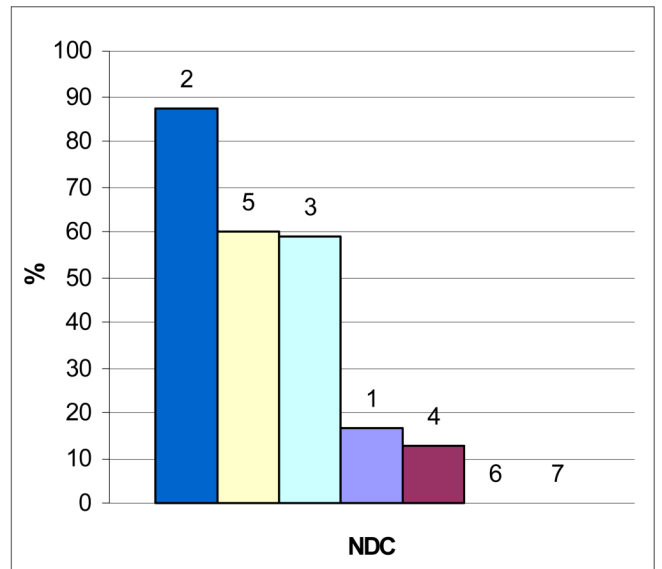
Figure 1. Sensitivity of automated clinical data for capturing chemotherapy administrations compared to tumor registries by health plan

Each bar represents a single health plan indicated by a color and number (1-7). Health plans are ordered from highest to lowest sensitivity for each type of automated clinical data. For example, health plan 1 has the highest sensitivity for HCPCS codes alone (90.2%), second highest for ICD-9-CM codes alone (65.9%), and fourth highest for NDCs alone (26.8%). When using all three types of codes, the overall sensitivity for health plan 1 is 92.7%.

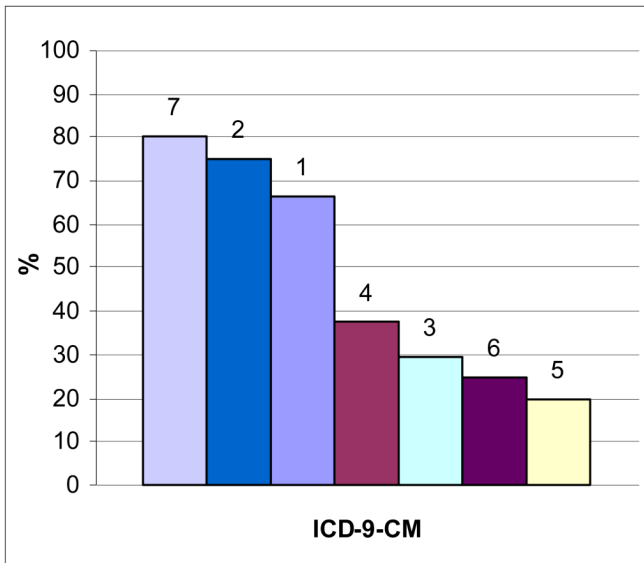
2a. Sensitivity of HCPCS codes



2b. Sensitivity of NDCs



2c. Sensitivity of ICD-9-CM codes



2d. Sensitivity of HCPCS + NDC + ICD-9-CM

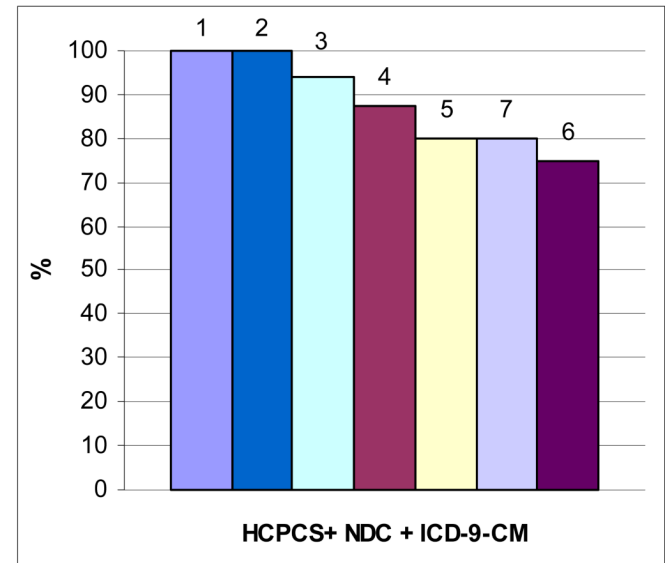
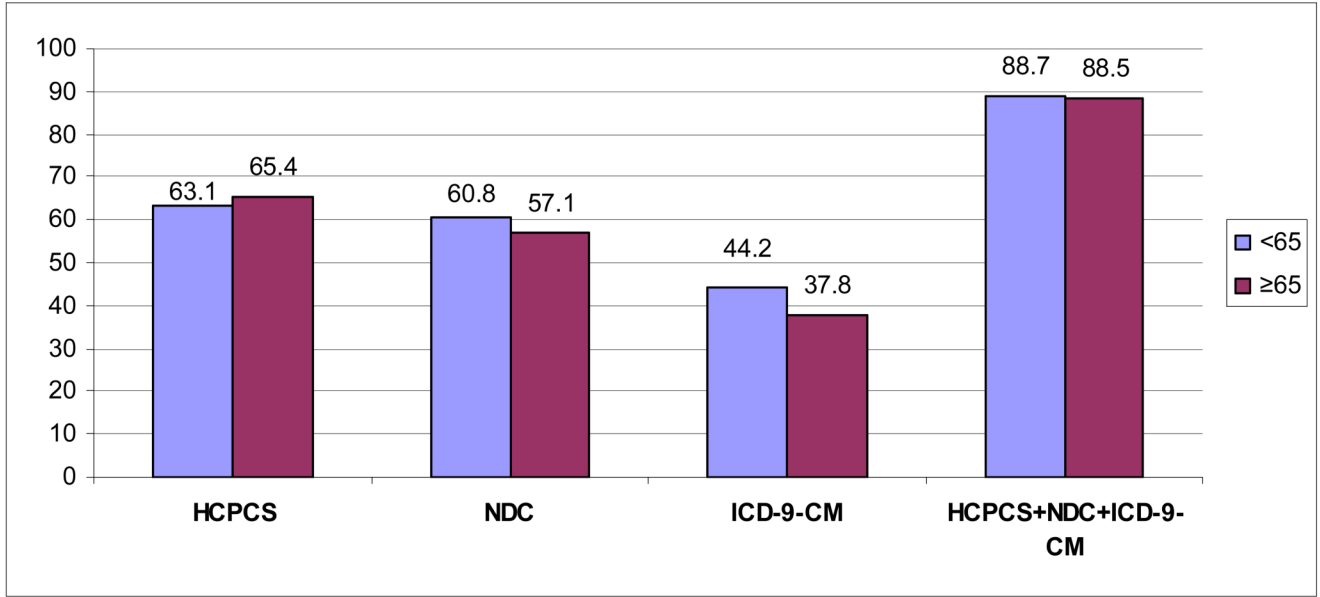


Figure 2. Sensitivity of automated clinical data for capturing chemotherapy administrations compared to medical records by health plan

Each bar represents a single health plan indicated by a color and number (1-7). Health plans are ordered from highest to lowest sensitivity for each type of automated clinical data. For example, health plan 1 has the highest sensitivity for HCPCS codes alone (100%), third highest for ICD-9-CM codes alone (66.7%), and fourth highest for NDCs alone (16.7%). When using all three types of codes, the sensitivity for health plan 1 is 100%.

3a. HCPCS+NDC+ICD-9-CM vs tumor registries



3b. HCPCS+NDC+ICD-9-CM vs medical records

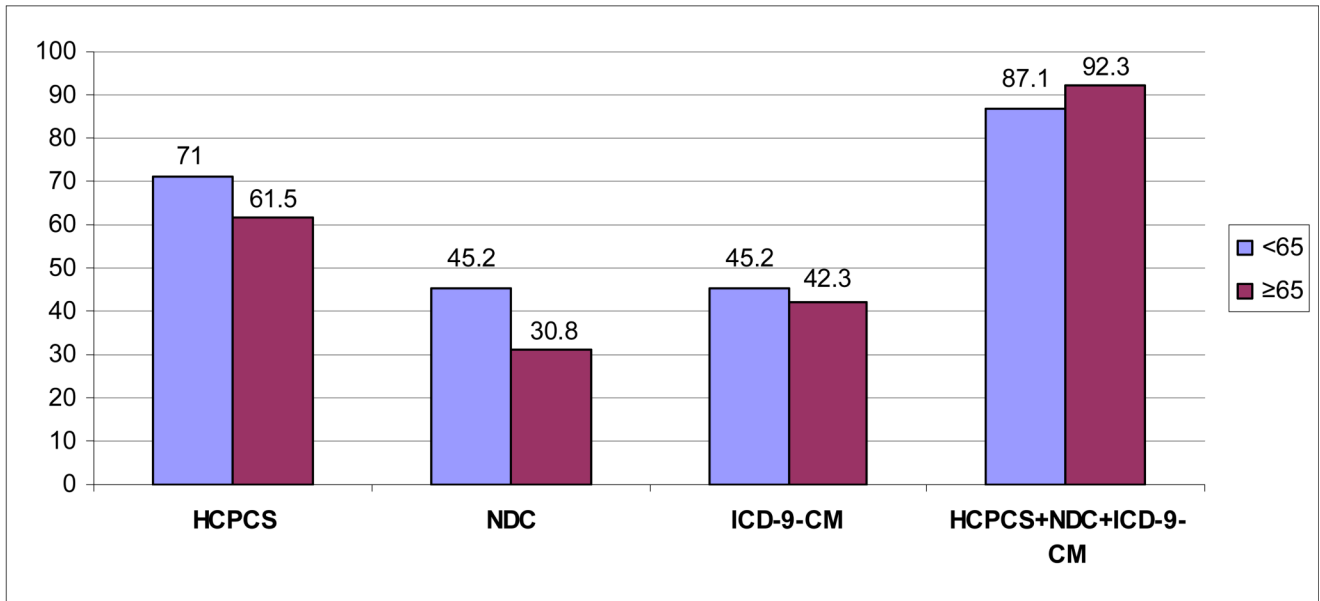


Figure 3. Sensitivity of automated clinical data for capturing chemotherapy administrations compared to tumor registries and medical records by age at diagnosis

Each bar represents the sensitivity for an age group (<65 or ≥65) indicated by a color with the sensitivity value at the top of the bar. For example, the sensitivity of HCPCS + NDC + ICD-9-CM codes combined for women <65 years was 88.7% compared to tumor registries (3a) and 87.1% compared to medical records (3b).

Table 1

Characteristics of 757 primary invasive ovarian cases including 70 chart reviewed cases diagnosed at CRN sites between Jan. 1, 2004-June 30, 2006¹

	n=757	All cases %	Subset of chart reviewed cases n=70 (%)
Age at diagnosis			
<30	11	(1.5)	0
30-39	28	(3.7)	<5
40-49	91	(12.0)	<5
50-59	195	(25.8)	16
60-69	188	(24.8)	26
70-79	145	(19.2)	17
80+	99	(13.1)	5
Race			
White	485	(69.6)	44
Black	26	(3.7)	5
Asian	78	(11.2)	<5
American Indian	<5	(0.1)	0
Other	<5	(0.6)	<5
More than one race	103	(14.8)	0
Missing	60		14
Hispanic ethnicity			
no	676	(94.2)	61
yes	42	(5.8)	<5
Missing	39		8
Year of diagnosis			
Jan.1, 2004-June 30, 2004	168	(22.2)	14
July 1, 2004-Dec.31,2004	160	(21.1)	15
Jan.1, 2005-June 30,2005	145	(19.2)	11
July 1, 2005-Dec.31,2005	147	(19.4)	10
Jan.1, 2006-June 30, 2006	137	(18.1)	20
Length of enrollment			
<5 years	254	(33.6)	22
5-<10 years	120	(15.9)	14
10-<15 years	105	(13.9)	10
15-<20 years	87	(11.5)	8
20+ years	191	(25.2)	16
Stage at diagnosis			
I	150	(21.2)	<5
II	63	(8.9)	0
III	300	(42.4)	41
IV	194	(27.4)	22
Missing	50		<5
Received surgery			
No	136	(18.2)	10
Yes	611	(81.8)	57
Missing	10		<5
Received chemotherapy (from tumor registries)			
No	171	(23.0)	9
Yes	572	(77.0)	60
Missing	14		<5
Morphology			
Serous	355	(46.9)	42
Mucinous	49	(6.5)	5
Papillary	14	(1.8)	0
Endometrioid	85	(11.2)	6
Clear cell	57	(7.5)	<5
Malignant/mixed mullerian	20	(2.6)	0
Other epithelial	174	(23.0)	14
Borderline	<5	(0.4)	0

¹ All data in this table are from automated clinical data, including tumor registry data.

Table 2 Sensitivity and specificity of automated clinical data for capturing chemotherapy procedures compared to tumor registries and medical records.

Automated clinical data codes	Cases with automated clinical data codes N (%)	Tumor registries (N=757)		Medical records (N=70)	
		Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
HCPCS only	378 (49.9)	64.0% (59.9, 67.9)	93.2% (88.5, 96.4)	66.7% (52.9, 78.6)	90.0% (55.5, 99.7)
NDC only	348 (46.0)	59.4% (55.3, 63.5)	95.5% (91.3, 98.0)	38.6% (26.0, 52.4)	100.0% (69.2, 100.0)
ICD-9-CM only	242 (32.0)	41.8% (37.7, 45.9)	98.3% (95.1, 99.6)	43.9% (30.7, 57.6)	100.0% (69.2, 100.0)
HCPCS + NDC	467 (61.7)	79.0% (75.5, 82.3)	91.5% (86.4, 95.2)	80.7% (68.1, 90.0)	90.0% (55.5, 99.7)
HCPCS + ICD-9-CM	468 (61.8)	79.7% (76.2, 82.9)	93.2% (88.5, 96.4)	87.7% (76.3, 94.9)	90.0 % (55.5, 99.7)
NDC + ICD-9-CM	457 (60.4)	78.1% (74.5, 81.5)	94.4% (89.9, 97.3)	63.2% (49.3, 75.6)	100.0% (69.2, 100.0)
HCPCS+ NDC + ICD-9-CM	522 (67.0)	88.6% (85.7, 91.1)	91.5% (86.4, 95.2)	89.5% (78.5, 96.0)	90.0% (55.5, 99.7)