Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data

Edmund C Lau[1]
Fionna S Mowat[1]
Michael A Kelsh[1,*]
Jason C Legg[2]
Nicole M Engel-Nitz[3]
Heather N Watson[1]
Helen L Collins[2]
Robert J Nordyke[4,5]
Joanna L Whyte[2]

[1]Exponent, Menlo Park, CA, USA; [2]Amgen, Thousand Oaks, CA, USA; [3]i3 Innovus, Eden Prairie, MN, USA; [4]PriceSpective LLC, El Segundo, CA, USA; [5]Department of Health Services, UCLA School of Public Health, Los Angeles, CA, USA
*Now at Amgen, Thousand Oaks, CA, USA

Abstract: Electronic medical records (EMRs) are used increasingly for research in clinical oncology, epidemiology, and comparative effectiveness research (CER).

Objective: To assess the utility of using EMR data in population-based cancer research by comparing a database of EMRs from community oncology clinics against Surveillance Epidemiology and End Results (SEER) cancer registry data and two claims databases (Medicare and commercial claims).

Study design and setting: Demographic, clinical, and treatment patterns in the EMR, SEER, Medicare, and commercial claims data were compared using six tumor sites: breast, lung/bronchus, head/neck, colorectal, prostate, and non-Hodgkin's lymphoma (NHL). We identified various challenges in data standardization and selection of appropriate statistical procedures. We describe the patient and clinic inclusion criteria, treatment definitions, and consideration of the administrative and clinical purposes of the EMR, registry, and claims data to address these challenges.

Results: Sex and 10-year age distributions of patient populations for each tumor site were generally similar across the data sets. We observed several differences in racial composition and treatment patterns, and modest differences in distribution of tumor site.

Conclusion: Our experience with an oncology EMR database identified several factors that must be considered when using EMRs for research purposes or generalizing results to the US cancer population. These factors were related primarily to evaluation of treatment patterns, including evaluation of stage, geographic location, race, and specialization of the medical facilities. While many specialty EMRs may not provide the breadth of data on medical care, as found in comprehensive claims databases and EMR systems, they can provide detailed clinical data not found in claims that are extremely important in conducting epidemiologic and outcomes research.

Keywords: electronic health records, data generalizability, oncology research, health care claims data, epidemiology

Introduction

EMRs are being used increasingly for observational research, post-marketing safety evaluation, and to inform decision making.[1–7] In February 2011, the Food and Drug Administration issued draft guidance regarding best practices for use of EMR in conducting pharmacoepidemiologic studies.[8] The primary advantages of EMRs include their potentially comprehensive and relatively timely clinical information, with the possibility of including physicians' notes, patient symptoms and history, diagnostic information, and planned and actual treatments. EMRs offer data not typically

Correspondence: Fionna Mowat
Exponent, 149 Commonwealth Drive, Menlo Park, CA, USA 94025
Tel +1 650 688 1782
Fax +1 650 688 1799
Email fmowat@exponent.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/CLEP.S23690

available in disease registries, claims records, or prescription databases, and are easier to analyze and usually more cost effective than chart reviews.

The availability of data varies across EMRs and depends on their design and completeness of data entry into applicable fields.[1] Because EMRs are designed primarily for patient care or billing, details that are important for health research (eg, tumor classification, unrelated comorbidities) may not be collected as rigorously as required for such research. Based on 2010 survey data, EMRs were adopted by approximately 50% of office-based physicians, increasing more than 30% from 2009; however, only 10% of EMR systems were termed fully functional.[9] Only 1.5% in 2008 and 2.7% in 2009 of the hospitals surveyed had comprehensive EMR systems, defined as including different levels of clinical functionality (eg, clinical documentation, test/image results) and decision support resources (eg, guidelines, drug alerts/interactions).[10–12] EMR adoption varies by state[9] and has occurred primarily in larger practices, urban areas, hospitals, or facilities owned by health maintenance organizations (HMOs).[10,11,13]

To evaluate the utility of EMRs in population-based cancer research, we compared demographic, clinical, and treatment factors from an aggregated US community oncology clinic EMR database with three other common data sources: Surveillance Epidemiology and End Results (SEER) registry, Medicare claims, and a large US commercial health insurance claims database. Our primary aim was to compare the content and utility of an oncology EMR vs claims data and a cancer registry and to better understand pertinent characteristics of oncology EMRs when using them to conduct epidemiologic and outcomes research studies. We also aimed to assess the benefits and results of applying data imputation procedures to missing data, to improve the completeness of data for comparison.

Methods

Data sources

The three comparison databases (SEER, Medicare, commercial claims) were chosen for their high-quality data and rigorous data collection and processing methods. Numerous published studies have relied on these sources. In addition, groups such as the National Cancer Policy Board, Institute of Medicine, and others have called for the strategic linking, cross-validation, and evaluation of registry and claims data with medical records to ensure quality care for oncology patients. Thus, these databases are considered relevant for evaluation of cancer research applications.[14–18] The most recent SEER and Medicare data available at the time of analysis were for 2006; therefore, 2006 data were selected for all data sources for comparative analyses.

EMR

The Oncology Services Comprehensive Electronic Records (OSCER) data warehouse is a proprietary database of EMRs from 52 outpatient oncology/hematology practice groups (15 hospital-affiliated and 37 community office-based) operating at 145 clinical sites that was initially formed by merging two EMR systems (Varian and IMPAC) and maintained by SDI Health. Data records contained each patient's diagnoses, clinic visits, and treatment, linked to visit dates. The integrated EMR warehouse has a single structure regardless of EMR source and can be analyzed in a single database. While the site practice groups varied in completeness of various data fields, the two source EMR systems were comparable in the fields available and used for analysis.

SEER

The National Cancer Institute's SEER program collects cancer incidence data from 17 population-based registries representing 26% of the US population.[19] SEER is considered a high-quality source for US cancer incidence, prevalence, histology, stage at diagnosis, and survival data. Except when compiled in special SEER studies (eg, Patterns of Care [POC]), prospective cancer treatment and clinical data are not collected from SEER-reported cancer patients.

Medicare

Medicare data, containing the claims history of about 2 million persons sampled from 35.2 million beneficiaries, provided a nationally representative source of medical treatment data for elderly US residents. Using the 5% sample of Medicare beneficiaries, we restricted the analysis to elderly patients (≥65 years old) who were diagnosed and treated for cancer. Beneficiaries included those receiving care through traditional pay-for-service programs; those enrolled in HMOs were excluded.[20] Diagnosis and treatment records were combined from inpatient (Part A), outpatient, and physician claims records (Part B); Part D records (pharmacy) were not included. Medical claims included those identified with diagnosis codes recorded with International Classification of Diseases (ICD-9-CM), and with procedural codes such as Current Procedural Terminology (CPT) and health care common procedure coding system (HCPCS) to identify patients of interest.

Commercial claims

Medical information from a commercial claims database was used to compare the demographic and treatment data among cancer patients < 65 years old. These data are derived from approximately 14 million employees and dependents across various employer-sponsored health plans and self-insured programs per year. Data include information on services from all available healthcare sites (in- and outpatient, emergency room, physician's office, surgery center) and all types of services (specialty, preventive, office-based treatments). The commercial claims records included diagnosis codes recorded with ICD-9-CM and recorded with CPT and HCPCS procedural codes, as well as site of service, provider specialty, and revenue data. Pharmacy claims data were not analyzed.

Patient selection

The specific patient selection criteria for selected tumor sites varied slightly for each data source (Table 1). Six sites of primary malignancies representing those most commonly reported in EMR and SEER were selected: breast, prostate, lung/bronchus, non-Hodgkin's lymphoma (NHL), colorectal (CRC), and head/neck tumors. Patients with all other forms of primary malignancy were grouped as "other tumors."

EMR

Patients were ≥20 years old; first seen at community outpatient oncology clinics, diagnosed with cancer between January 1 and December 31, 2006, and received ambulatory chemotherapy within 6 months of initial diagnosis (Table 1). To avoid classifying a patient as having cancer based exclusively on rule-out diagnoses, qualifying patients had at least two clinic visits within a 6-month period. Patients with diagnoses solely using V-codes (history of a particular cancer), tumors described as having uncertain behavior, or sarcoma histology codes were excluded (~7.3% of EMR records). The final analytical file comprised 169,199 patient records with 179,046 distinct primary tumors (Figure 1) over 7 years, with 31,117 patients and 32,357 distinct primary tumors in 2006. Certain clinics assigned an EMR loading date as a placeholder date, rather than the actual diagnosis or clinic visit date; these patient records were excluded.

SEER

Equivalent exclusion criteria used in the EMR data regarding age and diagnosis were applied to 2006 SEER data. Cancers identified only from death certificates or autopsy records and unconfirmed cancers (no or unknown microscopic or histological confirmation) were excluded, as were sarcomas (Table 1) per SEER recommendations.[21] Application of these criteria resulted in 20,780 patients excluded (~5.9%). Tumor sites of interest were identified using the SEER-recoded ICD-O-3 variable, which combines ICD-O-3 site and histology information.[19]

Medicare and commercial claims

Patients included those with two or more claims with the same cancer diagnosis in 2006 within 1 year in Medicare or at least 6 weeks apart in the commercial claims, and having no cancer diagnosis 12 months prior to the index diagnosis date (Table 1). To increase the likelihood of identifying new cancer diagnosis claims instead of prevalent conditions, patients were required to be covered continuously by the health plan or as a Medicare enrollee during the 12 months prior to the first cancer claim (ie, all Medicare patients were ≥66 years old). Claims only from laboratories, diagnostic testing centers, and diagnostic tests were not used to identify cancer claims, a procedure commonly used to avoid including false diagnoses.[22–24]

Factors evaluated

Patient demographic, clinical, and treatment characteristics were compared:

- Demographic: patient's age at diagnosis, sex, and race were compared across all four data sources.
- Clinical: tumor stage was compared in the EMR and SEER data. Tumor stage data were not available in the Medicare or commercial claims data.
- Treatment: ambulatory treatment was compared for EMR, Medicare, and commercial claims data. Ambulatory treatment was defined as cancer treatments provided in a clinical or office setting and divided into chemotherapy, biologics, and hormones. Chemotherapy was defined as cytotoxic intravenous drugs; biologics as small molecule-targeted therapies and anything produced by biotechnological/recombinant methods; and hormone therapies were limited to injectable medications administered in the clinic. Pharmacy-dispensed medications, investigational drugs, and radiology treatments were excluded.

Ambulatory treatment in the EMR was compared to similarly defined treatment in the Medicare (patients ≥ 65 years) and commercial claims (patients < 65 years) databases, stratified by tumor site, age, and sex of the patient. Because patients can be treated by more than one agent, the same patient could be counted in different treatment groups. To better standardize

Table 1 Description, characteristics, and inclusion/exclusion criteria of the four data sources

Source	Geographic regions covered	Identification of tumor sites	Identification of ambulatory treatment	Identification of patients/facilities
EMR	145 sites in 27 states	– ICD-9-CM – Excluded: • Tumors identified solely by V-codes • Codes indicating uncertain behavior • Sarcoma histology codes	– HCPCS – Ambulatory treatment received within 6 months of initial diagnosis	– Diagnosed with cancer between January 1, 2006 and December 31, 2006 – Age ≥ 20 years – At least two visits to a clinic within 6 months of cancer diagnosis – Excluded: • Facilities only administering radiation therapy • Clinics with administrative information only for diagnosis date
SEER	17 registries through the US	– ICD-O-3 – Included in situ or malignant tumors retained – Excluded patients with: • Unconfirmed diagnosis • Sarcoma histology codes	Not applicable	– Diagnosed with cancer between January 1, 2006 and December 31, 2006 – Age ≥ 20 years – No prior diagnosis of cancer – Excluded patients identified solely from death certificates or autopsy records
Medicare	National	– ICD-9-CM (Part A, Part B, outpatient records) – Excluded: • Tumors identified solely by V-codes • Claims from laboratories, diagnostic testing centers and diagnostic tests • Unspecified or benign tumors	– HCPCS – Limited to drugs administered via infusion, intravenously or by injection	– Diagnosed with cancer between January 1, 2006 and December 31, 2006 – Age ≥ 66 years – At least two claims with the same diagnosis within approximately 12 months – No prior diagnosis of cancer within 12 months
Commercial claims	National	– ICD-9-CM – Excluded claims from laboratories, diagnostic testing centers and diagnostic tests.	– HCPCS – Only one claim required for medications found on medical claims – Excluded solely implantable treatment devices – Exclude patient claims indicating chemotherapy administration, without drug specification	– Diagnosed with cancer between January 1, 2006 and December 31, 2006 – Age 20 ≥ age < 65 years – Covered by a health plan 12 months prior to first cancer claim – At least two claims with cancer diagnosis 42 days apart – Excluded claims with rule out codes – No prior use of ambulatory treatment

Abbreviations: HCPCS, health care common procedure coding system; ICD, International Classification of Diseases.

comparisons of ambulatory treatment usage rate, we focused on treatments received within 6 months of initial diagnosis.

Analytic and statistical methods

We calculated treated proportions of patients by dividing the number of patients receiving ambulatory therapy by the total number of patients with the specific cancer diagnosis. Given the large sample sizes from the databases evaluated, traditional tests of significance resulted in statistically significant findings, even for small absolute differences. Therefore, we focused on descriptive comparisons and used Cohen's *w* effect size (ES) with a pooled standard deviation to assess

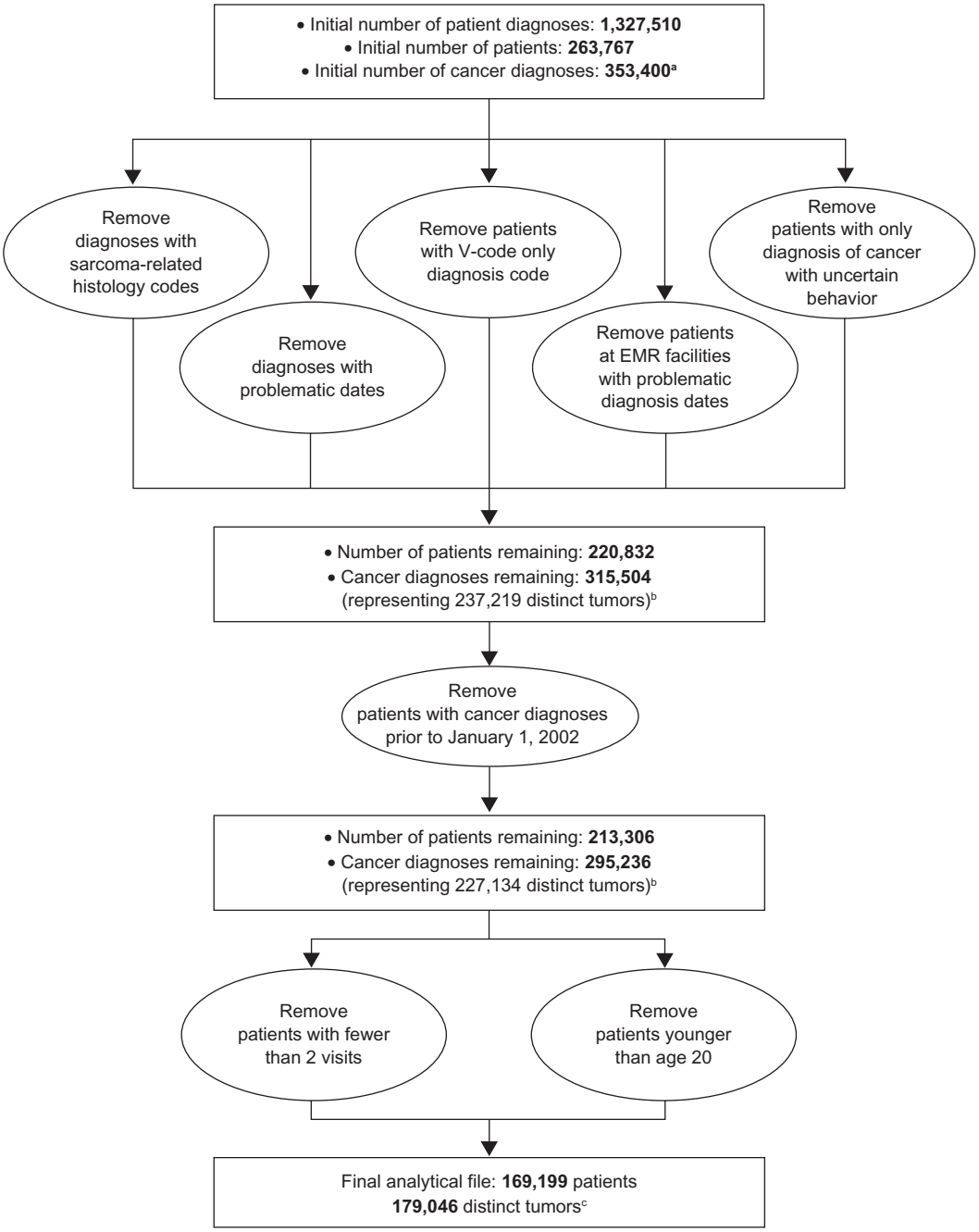

Figure 1 Schematic of processing and treatment of EMR oncology patient records.

Notes: For treatment analysis, EMRs without chemotherapy service and data were eliminated. Patients treated at oncology groups that did not provide chemotherapy services or those treated at radiology clinical sites within an oncology group were eliminated, resulting in a further reduced analytical file of 163,511 distinct tumors. [a]this was the number to which all subsequent exclusions were applied; [b]some patients had multiple tumors; [c]some patients had multiple tumors. This file includes some pre-2006 cancer patients that were not used.

Abbreviation: EMR, electronic medical records.

the importance of observed differences. This qualitative measure is not based on a rigorous hypothesis-testing framework and does not have a probabilistic interpretation such as a *P*-value obtained from standard methods. While the ES interpretation depends on the subject matter, Cohen classified the magnitude of the ES as small (w = ~0.1), medium (w = ~0.30), and large (w = ~0.50).[25,26]

A large proportion of data was missing for race (40%) and tumor stage (~70%) in the EMR records. The largest percentage of missing stage data (97%) was observed for NHL. Given that NHL treatment is determined mainly by subtype and pathology (not stage), this missing data trend was understandable. When these two categories were excluded, the proportion missing for stage was 63%. Text fields were not analyzed to determine whether they contained missing stage information. We selected a hot-deck method to impute missing data,[27] and compared this method against two other regression-based imputation procedures.[28–30] We also evaluated the model prediction properties of hot-deck imputation by applying it to records with known values. Sociodemographic information (2000 US Census) was incorporated into the imputation models. Pre- and post-imputation marginal distributions were compared to evaluate similarity in data sets and were found to be comparable to distributions of data among records with complete information for race and stage. An evaluation of the performance of the hot-deck procedure under a simple missing data mechanism that compared imputed and observed data was also conducted. Only post-imputed data comparisons are presented.

Results

SEER provided the largest number of patient records (331,427). There were 60,255 unique records in Medicare and 32,357 and 16,427 records in the EMR and commercial claims, respectively. Several differences were observed in overall tumor site distributions (Figure 2). Excluding the "other tumors" category, the largest proportion of patients had prostate cancer in Medicare, and the largest proportion of patients had breast cancer in the other three databases. In the oncology EMR data, >25% of the cancer patient records had breast tumors – nearly 7% more than the proportion in SEER – while Medicare had the lowest fraction (8%). The EMR had the highest percentage of lung cancer and NHL patients; proportions of patients with CRC or head/neck tumors were generally comparable across all databases. Prostate cancer was noticeably under-represented in the EMR, likely because prostate cancer patients are treated primarily by urologists.

Demographic characteristics

The mean ages of those ≥65 and <65 years in the EMR were 74.7 ± 6.6 years and 52.3 ± 9.2 years, respectively. Mean ages in Medicare (all ≥ 66 years) and commercial claims (all < 65 years) were 76.4 ± 7.2 and 51.4 ± 9.7 years, respectively. Age distributions were similar in EMR and SEER, except for 1) CRC, which had a younger age profile in EMR patients, and 2) prostate tumors, which had a larger proportion of ≥85-year old patients in the EMR (Table 2). Differences in sex distributions were generally <5%. Comparison of race was possible only for EMR

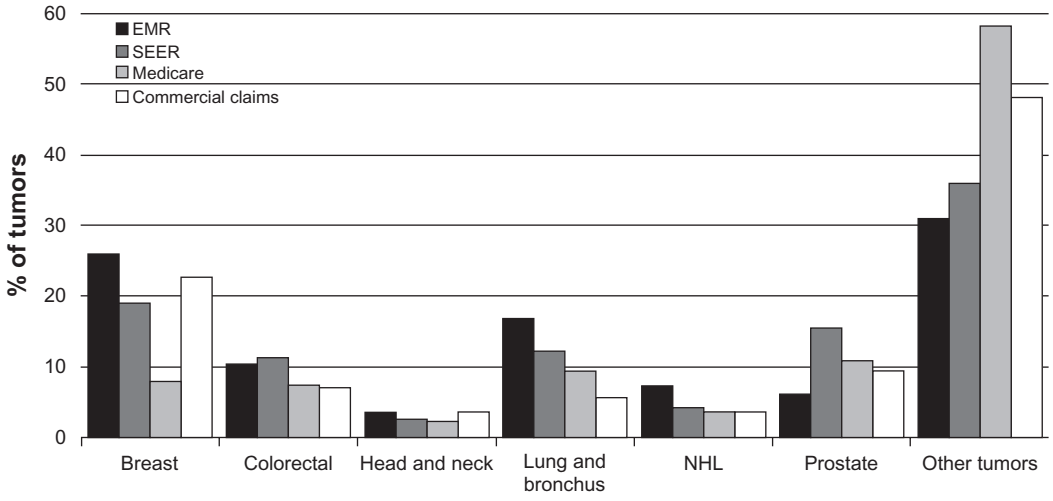

Figure 2 Distribution of tumor sites in oncology EMR and comparison databases.
Abbreviations: EMR, electronic medical records; SEER, Surveillance Epidemiology and End Results; NHL, non-Hodgkin's lymphoma.

Table 2 Age and gender distribution of oncology EMR and SEER patients by selected tumor site (2006 data)

Tumor site	Subgroup	# of qualifying patients in EMR	% by tumor site in EMR	# of qualifying patients in SEER	% by tumor site in SEER
Breast	**Female**	**8286**	**98.9**	**61,764**	**99.3**
	Male	**86**	**1.1**	**438**	**0.7**
	Age group (years)				
	20–34	151	1.8	986	1.6
	35–44	829	9.9	6603	10.7
	45–54	1990	23.9	14,955	24.1
	55–64	2214	26.4	15,479	24.9
	65–74	1769	21.1	12,105	19.4
	75–84	872	10.4	9178	14.7
	85+	547	6.5	2893	4.6
	Total	8372	100.0	62,202*	100.0
Colorectal	**Female**	**1585**	**47.7**	**18,406**	**49.3**
	Male	**1741**	**52.3**	**18,935**	**50.7**
	20–34	61	1.8	483	1.3
	35–44	183	5.5	1699	4.5
	45–54	532	16.0	5067	13.6
	55–64	851	25.6	7440	19.9
	65–74	870	26.2	9106	24.4
	75–84	504	15.2	9502	25.4
	85+	325	9.8	4042	10.8
	Total	3326	100.0	37,341*	100.0
Prostate	**Female**	**0**	**0**	**0**	**0**
	Male	**1,866**	**100**	**50,934**	**100**
	20–34	1	0.1	4	0.0
	35–44	3	0.2	345	0.7
	45–54	93	5.0	4671	9.2
	55–64	410	22.0	16,018	31.5
	65–74	625	33.5	18,138	35.6
	75–84	432	23.2	9977	19.6
	85+	302	16.2	1760	3.5
	Total	1866	100.0	50,934*	100.0
Head and neck	**Female**	**289**	**26.4**	**2668**	**32.4**
	Male	**804**	**73.6**	**5570**	**67.6**
	20–34	11	1.0	193	2.3
	35–44	67	6.1	546	6.6
	45–54	234	21.4	1768	21.5
	55–64	325	29.7	2290	27.8
	65–74	265	24.2	1742	21.1
	75–84	124	11.3	1256	15.2
	85+	67	6.1	443	5.4
	Total	1,093	100.0	8,238	100.0
Lung	**Female**	**2559**	**47.5**	**18,822**	**47.1**
	Male	**2830**	**52.5**	**21,180**	**52.9**
	20–34	25	0.5	110	0.3
	35–44	117	2.2	701	1.8
	45–54	631	11.7	3890	9.7
	55–64	1307	24.3	8950	22.4
	65–74	1853	34.4	12,858	32.1
	75–84	990	18.4	11,168	27.9
	85+	466	8.6	2325	5.8
	Total	5389	100.0	40,002	100.0

(*Continued*)

Table 2 (*Continued*)

Tumor site	Subgroup	# of qualifying patients in EMR	% by tumor site in EMR	# of qualifying patients in SEER	% by tumor site in SEER
NHL	**Female**	**1135**	**48.8**	**6494**	**46.8**
	Male	**1192**	**51.2**	**7395**	**53.2**
	20–34	104	4.5	510	3.7
	35–44	141	6.1	974	7.0
	45–54	331	14.2	1975	14.2
	55–64	514	22.1	2808	20.2
	65–74	588	25.3	3119	22.5
	75–84	366	15.7	3295	23.7
	85+	283	12.2	1207	8.7
	Total	2327	100.0	13,889*	100.0

Notes: *Several qualifying patients were missing age information in SEER. The totals shown include the patients with missing information. The number of patients with missing information is as follows: breast cancer, 3 patients; colorectal cancer, 2 patients; prostate cancer, 21 patients; and NHL, 1 patient.
Abbreviations: EMR, electronic medical records; SEER, Surveillance Epidemiology and End Results.

and SEER data; small differences ($w < 0.15$, all tumor sites) were observed. For each tumor site, the proportion of black patients in the EMR was higher than in SEER (Figure 3), likely because of geographic differences. The EMR groups were more concentrated in the South (41.2%) compared to SEER, which is more concentrated in western states (59.4%) and less in the South (10.1%) (Figure 4). Racial differences varied by tumor site (Figure 3).

Clinical characteristics

Comparison of tumor stage was possible only for EMR and SEER data. In SEER, stage is recorded at diagnosis, whereas stage in the EMR could be recorded either at diagnosis or at first visit. Comparison of the distributions of stage I and IV by tumor site in EMR and SEER indicated small differences ($w < 0.20$) (Figures 5 and 6). EMR and SEER had a similar percentage of stage I breast cancer patients, but SEER had higher proportions of stage I colorectal, head/neck, and lung/bronchus cancer than the EMR, likely because early stages of these diseases rarely require the systemic therapy offered by oncology clinics. The EMR data had a greater proportion of stage IV patients recorded for breast cancer, CRC, head/neck cancer, prostate cancer, and other cancers combined compared to SEER (Figure 6).

Treatment characteristics

Differences were observed in ambulatory treatment. In general, for all tumor sites except NHL and prostate cancer, a much larger percentage of patients in the EMR were treated compared to patients in the Medicare data (Table 3). This may be an artifact of age, where elderly patients with lower life expectancy might not be treated as aggressively as

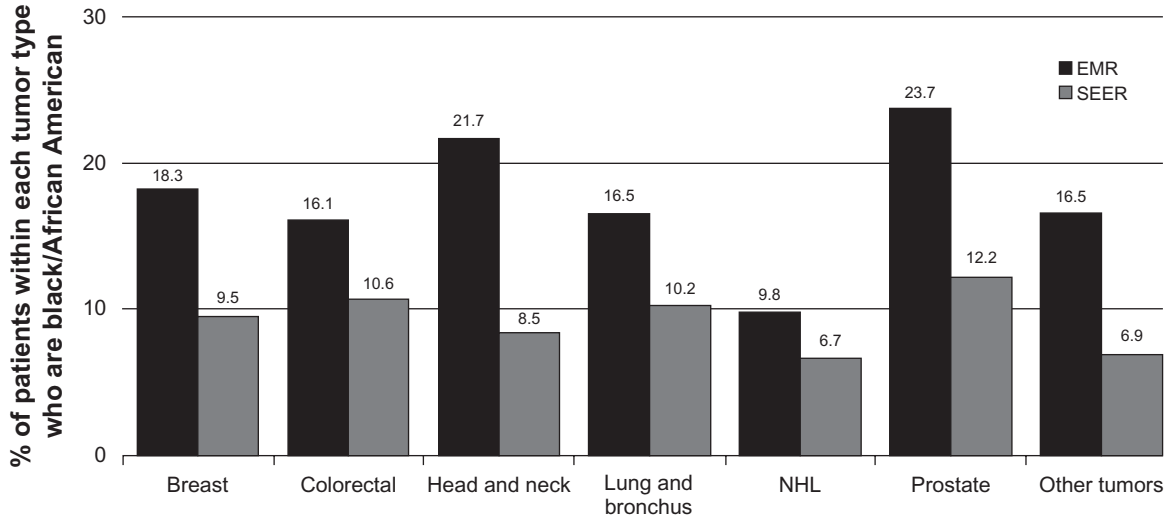

Figure 3 Percent black/African American cancer patients post-imputation by tumor site in oncology EMR and SEER databases.
Abbreviations: EMR, electronic medical records; SEER, Surveillance Epidemiology and End Results; NHL, non-Hodgkin's lymphoma.

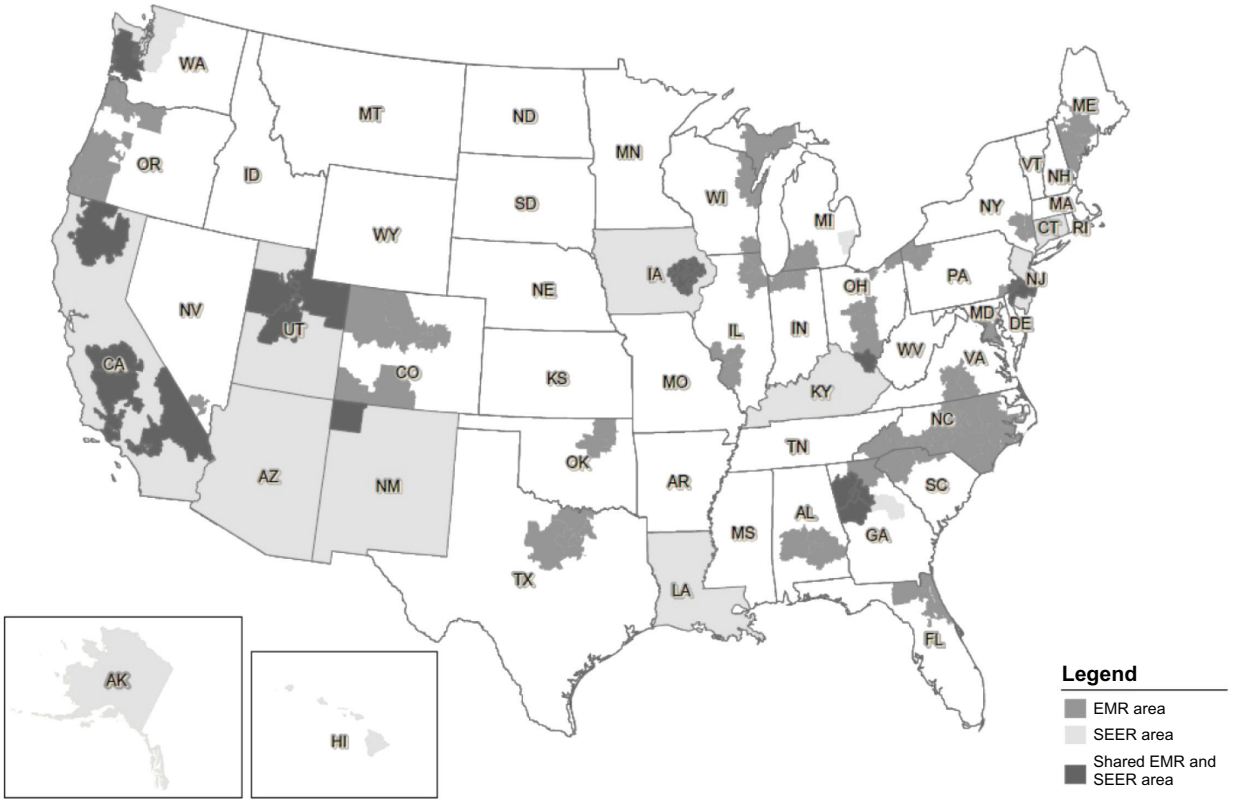

Figure 4 Geographic distribution of oncology EMR and SEER patients.
Abbreviations: EMR, electronic medical records; SEER, Surveillance Epidemiology and End Results.

other oncology patients. Among treated patients in the EMR and Medicare populations, EMR patients were more often treated with chemotherapy, except for breast cancer (Note: oral chemotherapy was not captured) and NHL. Patients in the EMR received more biologics and less hormone therapy than patients in the other databases, except for 1) lung cancer, where there were no differences in biologics, and 2) breast cancer, where hormones were more prevalent in breast cancer patients in the EMR (Table 4). Compared to Medicare, a much larger percentage of EMR patients received chemotherapy (44% vs 5%) among elderly prostate cancer patients. Among non-elderly patients (<65 years), the fraction of treated patients in the EMR was higher than in the commercial claims (Table 3). Differences in treatment

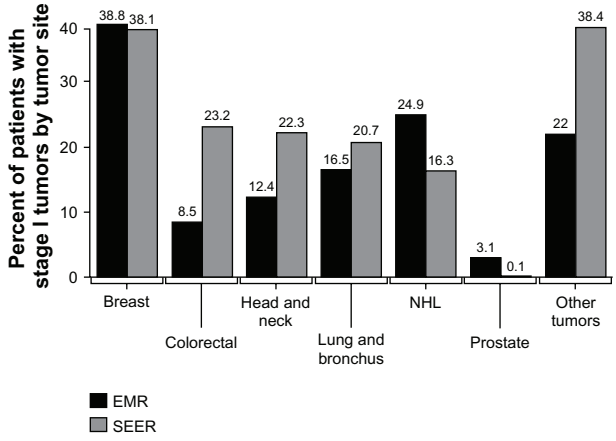

Figure 5 Post-imputation distribution of stage I tumors, by tumor site, at time of presentation in oncology EMR and at time of registration in SEER patients.
Abbreviations: EMR, electronic medical records; NHL, non-Hodgkin's lymphoma; SEER, Surveillance Epidemiology and End Results.

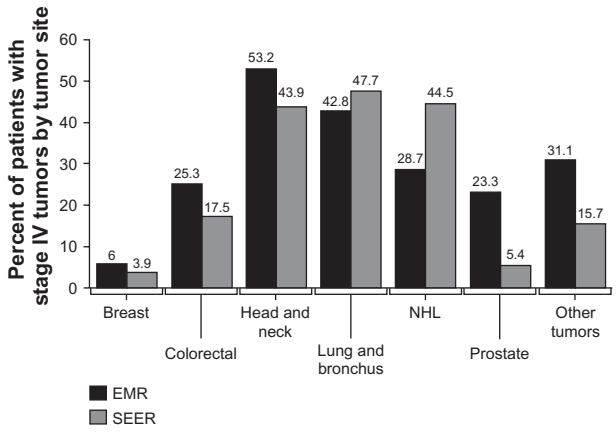

Figure 6 Post-imputation distribution of stage IV tumors, by tumor site, at time of presentation in oncology EMR and at time of registration in SEER patients.
Abbreviations: EMR, electronic medical records; NHL, non-Hodgkin's lymphoma; SEER, Surveillance Epidemiology and End Results.

Table 3 Percent of patients receiving outpatient hormone, biologic, or chemotherapy in oncology EMR, medicare, and commercial claims by tumor site and age category, 2006

	EMR patients (≥65 years)	Medicare	EMR vs Medicare	EMR patients (<65 years)	Commercial claims	EMR vs commercial claims
	% treated	% treated	% difference	% treated	% treated	% difference
Breast	20.6	13.7	6.9	44.4	28.1	16.3
Colorectal	40.7	20.1	20.5	60.4	45.8	14.7
Head/neck	51.1	21.0	30.0	55.1	22.1	33.0
Lung/bronchus	50.7	34.4	16.3	58.5	55.2	3.3
NHL	41.2	44.6	–3.4	47.6	38.4	9.2
Prostate	24.2	32.5	–8.3	23.1	12.8	10.3
Total[1]	34.8	16.1	18.6	45.9	23.7	22.3

Note: [1]Total includes other cancers and the six cancers listed above.
Abbreviations: EMR, electronic medical records; NHL, non-Hodgkin's lymphoma.

distribution among patients treated with chemotherapy and hormones were generally <5%, except in prostate cancer patients (Table 4). For all tumor sites except NHL, patients in the EMR were more likely to be treated with biologics than those in the commercial claims database.

Discussion

In this comparative analysis, similar distributions were observed in all four databases with respect to age and sex characteristics within specific tumor sites ($w < 0.3$); there were modest differences in racial composition (≤15%) and treatment patterns (≤30%) for several tumor sites. The overall tumor site distribution varied, with more breast cancer and proportionally fewer prostate cancer patients in the EMR. The greater percentage of breast cancer patients in the EMR compared to Medicare may be due to the high treatment rate of breast cancer patients in outpatient oncology clinics and because breast cancer occurs at a younger median age than other cancers. Thus, while breast cancer survivors certainly exist in the Medicare population, the number of newly diagnosed patients is expected to be small. Prostate cancer was noticeably under-represented in this oncology EMR, likely because these patients are treated primarily by urologists prior to referral to an oncologist, and treatment is less aggressive compared to other tumors. This limits the comparison of prostate cancer patients in the EMR with those in SEER and claims data. A large difference was observed among elderly prostate cancer patients in EMR compared to Medicare, most likely because most patients are referred to an oncologist after they have become hormone-refractory. Oncology patients in Medicare may also be less likely to be treated given their advanced age and decreased ability to tolerate aggressive chemotherapy.

Although the use of administrative data, claims-based databases, and other secondary data sources in epidemiologic research may not be ideal, they are well suited for certain types of analyses.[31,32] Other types of data commonly used in epidemiologic research, such as SEER data, are also well suited for certain analyses, but can be problematic in that they may be over- or under-representative of certain populations. In this analysis, our approach used a combination of three different databases, including a large commercial claims database, to circumvent database-specific issues and evaluate the utility of our EMR database from a broader perspective. In addition, we reported results in

Table 4 Percent of treated patients receiving hormone, biologic, or chemotherapy in oncology EMR, Medicare, and commercial claims databases patients by tumor site

Tumor site	EMR patients ≥ 65 years			Medicare			EMR patients < 65 years			Commercial claims		
	% C	% H	% B	% C	% H	% B	% C	% H	% B	% C	% H	% B
Breast	85.8	10.8	22.5	91.6	3.4	20.4	94.8	5.3	23.9	95.5	4.5	16.7
Colorectal	94.6	1.6	33.3	91.1	8.1	25.7	98.4	0.1	39.8	99.6	0.2	26.7
Head/neck	79.1	2.1	38.2	72.4	7.1	35.8	89.9	0.0	29.1	92.0	0.8	22.4
Lung/bronchus	98.4	0.6	11.3	97.1	3.2	11.3	99.5	0.5	14.6	99.8	0.4	10.8
NHL	72.2	0.2	83.9	79.6	1.8	78.3	85.8	0.6	74.0	88.9	0.4	80.8
Prostate	43.8	65.8	6.3	5.1	96	1.7	46.7	63.3	4.4	7.6	94.4	1.0
Total[1]	88.3	5.9	24	68.1	27.9	19.5	94.7	2.8	26.1	91.2	7.1	18.9

Note: [1]Total includes other cancers and the six cancers listed above.
Abbreviations: C, chemotherapy; H, hormones; B, biologics; EMR, electronic medical records; NHL, non-Hodgkin's lymphoma.

terms of relative differences (ie, effects sizes, proportions or percentages) rather than absolute differences to facilitate interpretation inference, as recommended by Sorensen et al.[33] The current analysis was focused solely on the US, mostly because the EMR data was collected from US oncology patients and our goal was to understand the generalizability of this database to the general US cancer population. While a similar approach could be used in different countries, EMR and claims-based systems will vary by country based on payment system; thus, this approach may or may not be applicable.

Many specific data comparability and methodological challenges arose during our analyses. These challenges fall into two general areas: 1) missing data and data standardization, and 2) patient/clinic characterization. For each of these challenges, we evaluated several approaches and assumptions and developed solutions to enable comparisons. Other challenges in using EMR and claims-based data are described elsewhere.[34,35] Data vocabulary issues were also challenging. For example, our classification of treatment as biologic, hormone, or chemotherapy was performed a priori, with input from oncologists and pharmacologists; however, some biologics may be classified differently by others.

Missing EMR data can introduce various problems in their validity in cancer research. In our analyses, the stage was frequently missing in the EMR and, though it was possibly recorded in the physician's notes or reports, was neither abstracted nor recorded in the EMR stage field. Patient race was also frequently missing in the EMR. Evaluation of text-based fields may result in additional data, particularly the stage, but was beyond the scope of this analysis, leading to the need for data imputation. Although imputation procedures were applied to fill data gaps to enable comparative analyses, the selection of an appropriate imputation method and subsequent validation of its application were challenging, especially because as much as 70% of the stage data was missing in some cases. This was considered in selecting the imputation method, and results were validated in a number of ways to reduce bias. The imputation methods allowed for population-level comparisons with other oncology data; however, at the individual record level, our validation study showed that the imputed data assignment may vary considerably from the true value, such that analyses that rely on individual-level imputed values will have errors due to incorrect classification.

Another challenge in data comparisons stemmed from patient selection, where patients from claims databases included all those who had at least two claims for any non-diagnosing visit with a cancer diagnosis either 6 weeks apart (commercial claims) or within 1 year (Medicare), independent of physician specialty and treatment type. In comparison, the EMR population included any patients with a cancer diagnosis and at least two visits in an oncology clinic, independent of treatment status. Since a patient seeing an oncologist is more likely to be treated, a much larger percentage of patients in the EMR received some type of therapy compared with the percentage in the claims databases, even when database tumor site distribution was similar. For example, for all tumor sites, except prostate cancer and NHL, in patients aged 65 or over, patients in the EMR were more frequently treated than those in other databases. The evaluation of treatment patterns warrants additional research, particularly because EMRs may aid in accelerating data collection and linking for evaluation of oncology treatment and patterns of care. Previous publications of treatment patterns[15–18] have been useful in measuring standards of care and changes over time.

In some cases, results obtained from EMR data were inconsistent with claims data, particularly treatment patterns. Differences in treatment patterns may be solely due to variance in tumor stage or to other factors, such as certain therapies being used more in the specialty oncology clinics. Differences in treatment could also be attributed simply to age and reduced life expectancy. From a clinical perspective, stage is usually the primary determinant of whether to treat (or not) a specific tumor with systemic therapy,[36] so the absence of stage data in the EMR fields and in commercial and Medicare claims is problematic to the interpretation of observed treatment differences. Treatment in EMR patients was not compared to SEER, since it was not available. Future studies might implement SEER POC studies,[37] which record information regarding cancer treatment reported in hospital and other medical records, to better understand treatment patterns in a national, population-based sample. Because outpatient treatment is often not well documented in hospital records, SEER POC studies[37] could aid in providing additional information on treatment among representative cancer patients. Similarly, the linked SEER-Medicare data files may present another source of data to better understand treatment and patient characteristics.

Other issues to consider in comparing findings across databases are selection factors that may influence treatments used and demographic profiles, because each database is derived from different populations, although some overlap exists. Racial differences may be explained by the fact that 40% of EMR facilities are in the southeastern region of the

US, with < 25% in the West, whereas 60% of SEER registries are in western states. Commercial claims and Medicare patients were included without any restriction applied to the type of clinic or physician visited. The included EMR facilities were referral oncology clinics, with patients who were more likely to require treatment by an oncologist, who might in turn be more likely to adopt the latest treatment technologies or use more aggressive treatment. Thus, treatment differences in the EMR and claims databases could be due to physician specialty or clinic and practice type.

There were several limitations in our analysis. We relied entirely on data provided in the EMR data warehouse; therefore, our analyses are limited to their accuracy and completeness and do not account for potential errors in data entry or misclassification. In addition, although we made every effort to isolate apparently newly diagnosed cancer cases in the EMR, due to previous clinic treatment not being recorded in this EMR, it is possible that some patients were being treated for recurrences of disease or had been treated before. This could also affect how the stage was recorded. Comparison of the proportion of stage IV patients in EMR and SEER data suggests that EMR patients have later stage cancer when appearing for care than the patients in SEER; however, this may be due to the fact that the stage is recorded at diagnosis in SEER and at a potentially unknown time period in disease progression for the EMR. Another limitation is the representativeness of our EMR data with other oncology data. The EMR data may differ from other databases due to patients' stage and disease severity and clinic/provider specialization; this was noted specifically for prostate cancer, where treatment by urologists (rather than oncologists) providing less aggressive treatment might lead to under-representation of this tumor in the EMR population. The fact that participating clinics were willing to provide and license their data for analysis may reflect other differences in these patient populations and treatment patterns when compared to other oncology databases. In addition, our analysis excluded pharmacy-dispensed medications and claims because they are not fully recorded in the EMR. Because some cancers are now treated with prescribed oral medications,[35,37] not necessarily recorded in the EMR, exclusion of pharmacy claims data may affect treatment interpretations. Finally, our focus on outpatient oncologic care may prevent comparison with other research that includes both inpatient and outpatient treatment.

The limitations of the data for purposes of this analysis relate to the timeliness of information and the availability of comprehensive comparable data fields across data sources. Our analysis was limited to 2006, because this was the most recent data available for SEER and Medicare at the time of analysis. Although EMRs were relatively novel in 2006, our entire EMR system was considered robust and included 263,767 patient records, 1,327,510 diagnosis records, 7,549,528 visits, and 3,828,337 records of the office-administered medications over 4 years. Medicare and commercial claims data represent administrative claims that are used primarily for insurance payment purposes and often lack important clinical information to help validate medical diagnosis and define disease progression. Given these limitations, various assumptions were made and operational definitions established to define and extract data for analysis. Although such assumptions are made frequently in studies that use administrative claims data, and could potentially be validated by merging claims and EMR, their validity was not assessed here. Finally, treatment percentages used may have been affected by the percentage of patients receiving active chemotherapy treatment vs follow-up and non-oncology clinic treatment, and may vary between EMR and reference patients.

Conclusions

Health care systems that use EMRs to track patients through medical service settings will offer a more complete, though sometimes still limited, source of data within that system. Our goals were to evaluate an oncology EMR database in comparison to a large cancer registry and two claims databases to characterize differences between the databases, and to subsequently use these comparisons to help in estimating characteristics in a broader target cancer population in the US. Our experience with an oncology EMR database identified several factors, including the stage, geographic location, and specialization of the medical facilities, that must be considered when using EMRs for research purposes or generalizing results to the US cancer population and for conducting epidemiologic research in general. While specialty EMRs may not provide the breadth of data on medical care, as found in comprehensive claims databases and EMR systems, specialty EMRs can provide detailed clinical data not found in claims that are extremely important in conducting epidemiologic and outcomes research.

Acknowledgments

The authors thank Jon Fryzek and Cynthia O'Malley for their input regarding the epidemiology of various tumor sites and aid in the initial stages of this project. The authors also thank Duane Steffey for his aid in developing and applying imputation procedures, and SDI Health LLC for provision

of the de-identified EMR data. Specific staff at SDI also aided in the data extraction and analyses, including Steve Lamond, Dawn Richardson, Susan Dennis, Robert Swift, and Melissa Pirolli. At i3 Innovus, Jane Sullivan programmed the analytic data set, and Gabriel Gomez Rey performed the statistical analyses.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Use of electronic medical records for health outcomes research: A literature review. *Med Care Res Rev*. 2010;66(6):611–638.
2. Kanas G, Morimoto L, Mowat F, et al. Use of electronic medical records in outcomes research. *ClinicoEconomics and Outcomes Res*. 2010; 2:1–14.
3. Bush GW. Exec. Order No. 13335: Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator Weekly Compilation of Presidential Documents 2004:702–704. Available at: http://www.gpo.gov/fdsys/pkg/WCPD-2004-05-03/pdf/WCPD-2004-05-03-Pg702.pdf. Accessed on September 8, 2011.
4. US Department of Health and Human Services (DHHS). Achieving a transformed and modernized health care system for the 21st Century: CMS strategic action plan, 2006–2009. October 16, 2006. Available at: http://www2.ancor.org/issues/medicaid/cms_strat_plan_06-09_10-06.pdf. Accessed on July 2, 2011.
5. US Food and Drug Administration (FDA). The Sentinel Initiative. National Strategy for Monitoring Medical Product Safety. Department of Health and Human Services (DHHS). May 2008. Available at: http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm. Accessed on July 2, 2011.
6. American Recovery and Reinvestment Act. 111th Congress (2009–2010) H.R.1.ENR. Division A – Appropriations Provisions, Title VIII – Departments of Labor, Health, and Human Services, and Education, and related agencies. 2009. Available at: http://thomas.loc.gov/cgi-bin/query/z?c111:H.R.1.enr:. Accessed on July 2, 2011.
7. Brookings Institution. Implementing comparative effectiveness research: Priorities, methods, and impact. 111th Congress of the United States of America. American Recovery and Reinvestment Act. Division A, Title VIII. Engelberg Center for Health Care Reform at Brookings. Washington, DC. June 9, 2009. Available at: http://www.brookings.edu/events/2009/0609_health_care_cer.aspx. Accessed June 16, 2010.
8. US Department of Health and Human Services (DHHS). Guidance for industry and FDA staff. Best practices for conducting and reporting pharmacoepidemiologic safety studies using electronic healthcare data sets. DHHS, US Food and Drug Administration (FDA), Center for Drugs Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). February 2011. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM243537.pdf. Accessed on July 2, 2011.
9. Hsiao CJ, Beatty PC, Hing ES, et al. Electronic medical record/electronic health record system of office-based physicians: United States, 2009 and preliminary 2010 state estimates. National Center for Health Statistics (NCHS) Health and Stats. NCHS, Centers for Disease Control and Prevention. December 2010. Available from: http://www.cdc.gov/nchs/data/hestat/emr_ehr_09/emr_ehr_09.htm. Accessed on August 29, 2011.
10. Jha AK, Ferris TG, Donelan K, et al. How common are electronic health records in the United States? A summary of the evidence. *Health Aff (Millwood)*. 2006;25(6):w496–w507.
11. Jha AK, DesRoches CM, Kralovec PD, Joshi MS. A progress report on electronic health records in U.S. Hospitals. *Health Aff (Millwood)*. 2010;29(10):1951–1957.
12. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in U.S. hospitals. *N Eng J Med*. 2009;360(16):1628–1638.
13. DesRoches CM, Campbell EG, Rao SR, et al. Electronic records in ambulatory care – A national survey of physicians. *N Engl J Med*. 2008;359(1):50–60.
14. Chassin MR, Gavin RW. The urgent need to improve health care quality. Institute of Medicine National Roundtable on Health Care Quality. *JAMA*. 1998;280(11):1000–1005.
15. Hillner BE, McDonald MK, Desch CE, Smith TJ, Penberthy LT, Retchin SM. A comparison of patterns of care of nonsmall cell lung carcinoma patients in a younger and Medigap commercially insured cohort. *Cancer*. 1998;83(9):1930–1937.
16. Hillner BE, McDonald MK, Penberthy L, et al. Measuring standards of care for early breast cancer in an insured population. *J Clin Oncol*. 1997;15(4):1401–1408.
17. Hewitt M, Simone JV, editors. *Ensuring Quality Cancer Care*. Washington, DC: National Academy Press; 1999.
18. Institute of Medicine (IOM). *Crossing the Quality Chasm*: *A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001.
19. National Cancer Institute (NCI). Surveillance Epidemiology and End Results (SEER). Population characteristics. US National Institutes of Health, National Cancer Institute 2010. Available at: http://seer.cancer.gov/registries/characteristics.html. Accessed June 16, 2010.
20. Centers for Medicare and Medicaid Services (CMS). Medicare enrollment reports. US Department of Health and Human Services (DHHS), Centers for Medicare and Medicaid Services, 2010. Available at: http://www.cms.hhs.gov/MedicareEnrpts/. Accessed June 16, 2010.
21. Ries LAG, Young JL, Jr., Keel GE, et al. (eds). SEER survival monograph: Cancer survival among adults: U. SEER Program, 1988–2001, Patient and tumor characteristics. National Cancer Institute, SEER Program, NIH Pub. No.07–6215, Bethesda, MD. 2007. Available at: http://seer.cancer.gov/publications/survival/seer_survival_mono_lowres.pdf. Accessed on July 2, 2011.
22. Duh MH, Reynolds WJ, Lefebvre P, Neary M, Skarin AT. Costs associated with intravenous chemotherapy administration in patients with small cell lung cancer: a retrospective claims database analysis. *Curr Med Res Opin*. 2008;24(4):967–974.
23. Lage MJ, Barber BL, Harrison DJ, Jun S. The cost of treating skeletal-related events in patients with prostate cancer. *Am J Manag Care*. 2008;14(5):317–322.
24. Ramsey SD, Martins RG, Blough DK, Tock LS, Lubeck D, Reyes CM. Second-line and third-line chemotherapy for lung cancer: use and cost. *Am J Manag Care*. 2008;14(5):297–306.
25. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, revised edition. New York: Academic Press; 1977.
26. Hayes WL. *Statistics*, 4th edition. Orlando, FL: Holt, Rinehart and Winston; 1988.
27. Altmayer L. Hot-Deck Imputation: A Simple DATA Step Approach. U.S. Bureau of the Census, Washington, DC, 1999. Available at: http://analytics.ncsu.edu/sesug/1999/075.pdf. Accessed on July 2, 2011.
28. Raghunathan TE, Solenberger PW, Van Hoewyk J. IVEware: Imputation and variance estimation software user guide. University of Michigan. 2002. Available at: http://www.isr.umich.edu/src/smp/ive/. Accessed on July 2, 2011.
29. Schafer JL. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall; 1997.
30. Allison PD. *Missing Data*. (Quantitative Applications in the Social Sciences, 07–136.) Thousand Oaks, CA: Sage Publications; 2002.
31. Grimes DA. Epidemiologic research using administrative databases: Garbage in, garbage out. *Obstet Gynecol*. 2010;116(5):1018–1019.
32. Hoover KW, Tao G, Kent CK, Aral SO. Letter to the editor. Epidemiologic research using administrative databases: Garbage in, garbage out. *Obstet Gynecol*. 2011;117(3):729.

33. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol*. 1996;25(2):435–442.
34. Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: Issues and challenges. *J Am Med Inform Soc*. 2010;17(6):671–674.
35. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323–337.
36. National Comprehensive Cancer Network (NCCN). Available at: http://www.NCCN.org. Accessed July 25, 2011.
37. Surveillance Epidemiology and End Results (SEER). Patterns of care / quality of care studies, 2010. Available at: http://healthservices.cancer.gov/surveys/poc. Accessed January 21, 2010.