Validation of Claims Data Algorithms to Identify Nonmelanoma Skin Cancer

Melody J. Eide[1,2], J. Mark Tuthill[3], Richard J. Krajenta[2], Gordon R. Jacobsen[2], Marc Levine[3] and Christine C. Johnson[2]

Health maintenance organization (HMO) administrative databases have been used as sampling frames for ascertaining nonmelanoma skin cancer (NMSC). However, because of the lack of tumor registry information on these cancers, these ascertainment methods have not been previously validated. NMSC cases arising from patients served by a staff model medical group and diagnosed between 1 January 2007 and 31 December 2008 were identified from claims data using three ascertainment strategies. These claims data cases were then compared with NMSC identified using natural language processing (NLP) of electronic pathology reports (EPRs), and sensitivity, specificity, positive and negative predictive values were calculated. Comparison of claims data–ascertained cases with the NLP demonstrated sensitivities ranging from 48 to 65% and specificities from 85 to 98%, with ICD-9-CM ascertainment demonstrating the highest case sensitivity, although the lowest specificity. HMO health plan claims data had a higher specificity than all-payer claims data. A comparison of EPR and clinic log registry cases showed a sensitivity of 98% and a specificity of 99%. Validation of administrative data to ascertain NMSC demonstrates respectable sensitivity and specificity, although NLP ascertainment was superior. There is a substantial difference in cases identified by NLP compared with claims data, suggesting that formal surveillance efforts should be considered.

Journal of Investigative Dermatology (2012) **132,** 2005–2009; doi:10.1038/jid.2012.98; published online 5 April 2012

INTRODUCTION

Skin cancer is becoming an increasing health burden (Athas *et al.*, 2003; Housman *et al.*, 2003; Rogers *et al.*, 2006). The majority of these skin cancers are basal cell carcinoma (BCC) and cutaneous squamous cell carcinoma (SCC), which are commonly referred to collectively as nonmelanoma skin cancer (NMSC), and represent the most common malignancy in the United States. Annual incidence of NMSC is estimated to be nearly equal to the incidence of all other cancers combined (Housman *et al.* 2003; Jemal *et al.*, 2010).

Previously, we defined and compared algorithms for identifying NMSC using the computerized administrative claims–based data set of a large health-care system provider and its affiliated health maintenance organization (HMO; Eide *et al.*, 2010). Using chart review of claims data algorithms examining International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and current procedural terminology codes (CPT), we demonstrated positive predictive values (PPVs) ranging from 47% for ICD-9-CM-ascertained cases to 95% for cases ascertained with both ICD-9-CM and CPT codes in a random sample of all-payer cases. NMSC cases were confirmed in >97% of cases regardless of the ascertained method in a sample of health-plan enrollees. The lack of tumor registry information, as these cancers are excluded from common tumor sources, including the Surveillance Epidemiology and End Results (SEER) program, prohibited validation of the algorithms against a gold-standard measure and an estimation of missed true cases.

Validation of claims data algorithms for NMSC ascertainment, including information on missed cases in claims data, is paramount for standardizing the study of NMSC. Capitalizing on health-system electronic pathology information, which is integral to the e-surveillance of reportable tumors to the local (SEER) tumor registry, we proposed to ascertain NMSC cases similarly from electronic pathology reports (EPRs). These electronic histopathology records would then constitute a gold-standard comparison for cases ascertained by claims data. The objective of this study was to determine sensitivity, specificity, PPV, and negative predictive value (NPV) of claims data algorithms to ascertain NMSC cases, with validation against the health-system EPR.

[1]*Department of Dermatology, Henry Ford Hospital, Detroit, Michigan, USA;* [2]*Department of Public Health Sciences, Henry Ford Hospital, Detroit, Michigan, USA and* [3]*Department of Pathology, Henry Ford Hospital, Detroit, Michigan, USA*

Correspondence: Melody J. Eide, Department of Dermatology, Henry Ford Hospital, 3031 West Grand Boulevard, Suite 800, Detroit, Michigan 48202, USA. E-mail: meide1@hfhs.org

Abbreviations: BCC, basal cell carcinoma; CPT, current procedural terminology code; EPR, electronic pathology report; HMO, health maintenance organization; ICD-9-CM, International Classification of Disease, Ninth Revision, Clinical Modification; NLP, natural language processing; NPV, negative predictive value; NMSC, nonmelanoma skin cancer; PPV, positive predictive value; SCC, squamous cell carcinoma; SEER, Surveillance Epidemiology and End Results

Received 14 December 2011; revised 7 February 2012; accepted 17 February 2012; published online 5 April 2012

RESULTS

From 1 January 2007 to 31 December 2008, there were 24,164 cases involving skin specimens processed by histopathology as identified by the EPR in the all-payer population. This included 4,883 unique NMSC cases. Comparison of all-payer claims data NMSC ascertainment algorithms to the EPR demonstrated sensitivities ranging from 48 to 64% and specificities from 85 to 94%, with ICD-9-CM ascertainment demonstrating the highest case sensitivity and the combination of ICD-9-CM and CPT together obtaining the highest specificity and PPV (Table 1).

In the HMO population, there were 15,297 total skin specimen cases and 2,506 cases of NMSC ascertained from claims data. When compared with the EPR, NMSC claims data algorithm sensitivities ranged from 49 to 65% and specificities from 96 to 98% (Table 2).

One clinic log site, which the EPR was cross-validated against, submitted 4,614 total cutaneous cases during the study period. This included 909 NMSC cases. The sensitivity, specificity, and negative and PPVs for the clinic logbook site and the EPR compared favorably, with all values >98% (Table 3). The logbook clinic site, all-payer claims data was comparable by NMSC ascertainment algorithm with the entire health-system estimates (Table 4).

The reasons for discordance of cases ascertained from the clinic logbook and the EPR were investigated. The majority of EPR non-confirmed cases or false cases was attributable to exclusion of residual cutaneous malignancy, likely on re-excision (Table 5).

DISCUSSION

We present validation of previously defined claims data NMSC ascertainment algorithms using the computerized databases compared with the results from the use of natural language processing (NLP) of electronic histopathology records of a large health system. Cases of NMSC can be ascertained in the health-system setting using administrative data with respectable sensitivity, specificity, NPVs and PPVs, with higher case sensitivity using ICD-9-CM ascertainment methods and higher specificity using both ICD-9-CM and CPT together, which we hope will provide interested investigators knowledge and direction as they design secondary data studies. Our findings also further the understanding of the capacity and limitations of using claims data to identify and investigate NMSC.

The World Health Organization recognizes the difficulty in ascertaining the incidence of NMSC, noting limited registries capturing BCC and SCC, especially in North

Table 1. Identified and confirmed nonmelanoma skin cancer cases with sensitivity, specificity, NPV, and PPV (2007–2008), all-payer claims data

Identification algorithm	True positive cases	Identified cases by algorithm	Confirmed true cases identified by algorithm	Sensitivity (95% CI)	Specificity (95% CI)	NPV (95% CI)	PPV (95% CI)
ICD-9-CM code alone	4,883	5,995	3,128	64.1 (62.7–65.4)	85.1 (84.6–85.6)	90.3 (89.9–90.8)	52.2 (50.9–53.4)
CPT code alone	4,883	5,541	3,078	63.0 (61.7–64.4)	87.2 (86.8–87.7)	90.3 (89.9–90.7)	55.5 (54.2–56.9)
Both ICD-9-CM and CPT codes	4,883	3,441	2,335	47.8 (46.4–49.2)	94.3 (93.9–94.6)	87.7 (87.3–88.2)	67.9 (66.3–69.4)

Abbreviations: CI, confidence interval; CPT, current procedural terminology; ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; NPV, negative predictive value; PPV, positive predictive value.
Study population: large integrated health system, all-payer health-plan patients, southeastern MI, USA.
Standard: Co-Path electronic histopathology data.
Total skin pathology cases: 24,164.

Table 2. Identified and confirmed nonmelanoma skin cancer cases with sensitivity, specificity, NPV, and PPV by algorithm (2007–2008), HMO-enrollee claims data

Identification algorithm	True positive cases	Identified cases by algorithm	Confirmed true cases identified by algorithm	Sensitivity (95% CI)	Specificity (95% CI)	NPV (95% CI)	PPV (95% CI)
ICD-9-CM code alone	2,506	2,209	1,639	65.4 (63.5–67.3)	95.5 (95.2–95.9)	93.4 (93.0–93.8)	74.2 (72.4–76.0)
CPT code alone	2,506	2,059	1,610	64.2 (62.4–66.1)	96.5 (96.2–96.8)	93.2 (92.8–93.7)	78.2 (76.4–80.0)
Both ICD-9-CM and CPT codes	2,506	1,534	1,230	49.1 (47.1–51.0)	97.6 (97.4–97.9)	90.7 (90.2–91.2)	80.2 (78.2–82.2)

Abbreviations: CI, confidence interval; CPT, current procedural terminology; HMO, health maintenance organization; ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; NPV, negative predictive value; PPV, positive predictive value.
Study population: large integrated health system, HMO health-plan enrollees, southeastern MI, USA.
Standard: Co-Path electronic histopathology data.
Total skin pathology cases: 15,297.

Table 3. Confirmed and identified nonmelanoma skin cancer cases with sensitivity, specificity, NPV, and PPV (2007–2008), electronic pathology record and logbook clinic site

Standard source	Confirmed cases by chart review	Identified by method	Sensitivity (95% CI)	Specificity (95% CI)	NPV (95% CI)	PPV (95% CI)
EPR	894	909	98.3 (97.5–99.2)	99.6 (99.4–99.8)	99.6 (99.4–99.8)	98.2 (97.4–99.1)
Clinic logbook	894	910	98.2 (97.4–99.1)	99.5 (99.3–99.7)	99.6 (99.4–99.8)	98.0 (97.1–98.9)

Abbreviations: CI, confidence interval; EPR, electronic histopathology record; NPV, negative predictive value; PPV, positive predictive value.
Study population: large integrated health system, all-payer health-plan patients, logbook clinic site, southeastern MI, USA with Co-Path EPR.
Total skin pathology cases: 4,614.
Standard: chart review of all identified cases regardless of ascertainment method.

Table 4. Total pathology skin specimens, identified and confirmed nonmelanoma skin cancer cases with sensitivity, specificity, NPV, and PPV, 2007–2008, clinic log data

Identification algorithm	Identified cases by algorithm	Confirmed cases	Sensitivity (95% CI)	Specificity (95% CI)	NPV (95% CI)	PPV (95% CI)
ICD-9-CM code alone	626	540	59.4 (56.2–62.6)	97.7 (97.2–98.2)	90.7 (89.8–91.6)	86.3 (83.6–89.0)
CPT code alone	544	501	55.1 (51.9–58.3)	98.8 (98.5–99.2)	90.0 (89.0–90.9)	92.1 (89.8–94.4)
Both ICD-9-CM and CPT codes	487	462	50.8 (47.6–54.1)	99.3 (99.1–99.6)	89.2 (88.2–90.1)	94.9 (92.9–96.8)

Abbreviations: CI, confidence interval; CPT, current procedural terminology; ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; NPV, negative predictive value; PPV, positive predictive value.
Study population: large integrated health system, single clinical site, southeastern MI, USA.
Standard: Co-Path electronic histopathology data.
Total skin pathology cases: 4,614.

Table 5. Characteristics of false cases of NMSC identified by either clinic log or EPR information (2007–2008)

	Unconfirmed cases	
Reason	EPR (*N*=15)	Clinic log (*N*=16)
Missing information (omission or abstractor error possible)	5	0
Basal cell or squamous cell carcinoma[1]	2	11
Rare cutaneous carcinomas not included in electronic search (e.g., ''sebaceous carcinoma'', ''desmoplastic epithelial tumor'')	0	2
Suggestion or inconclusive description of cutaneous malignancy (e.g., ''basaloid epitheloid islands'', ''suggestive of'')	1	2
Exclusion of cutaneous malignancy (e.g., ''scar, negative for'', ''negative for'', ''no residual'')	6	0
Different physical clinical location than clinic log site itself[1]	1	1

Abbreviations: EPR, electronic histopathology record; NMSC, nonmelanoma skin cancer.
[1]Patient was flagged as present in both columns due to nonmatching date of service differences (*N*=2; of which one had different clinic site and one was a squamous cell carcinoma).
Study population: large integrated health system, single clinical site and Co-Path electronic histopathology record data, southeastern MI, USA.

America (International Agency For Research On Cancer, 2008). It is a significant investment of time and resources to initiate new or expanded traditional registries for the ascertainment of BCC and SCC (Lamberg *et al.*, 2010). We believe that collaborative efforts of large US HMOs, such as those that participate in the National Cancer Institute funded Cancer Research Network (CRN), which covers nearly 11 million individuals, have the potential to provide high-quality, efficient NMSC ascertainment in the United States (National Cancer Institute, 2010).

We previously reported algorithms for identifying NMSC using the computerized administrative claims–based data set of this same large US health-care system provider and its affiliated HMO (Eide *et al.*, 2010). NMSC patients who were diagnosed between 1988 and 2007 were identified using three algorithms: NMSC ICD-9-CM codes, NMSC treatment CPT codes, or both ICD-9-CM and CPT codes. A subset of charts was reviewed to verify NMSC diagnosis, including all HMO enrollee members' EMRs in 2007, and PPV for NMSC were calculated (with sensitivity, specificity, and NPV unable to be predicted). A random sample of all years, and all payers

were selected for chart review, with PPVs of 47.0% for ICD-9-CM-identified patients, 73.4% for CPT-identified patients, and 94.9% identified with both codes required. All charts from HMO health-plan enrollees in 2007 were reviewed with PPVs of 96.5% for ICD-9-CM-identified patients, 98.3% for CPT-identified patients, and 98.7% identified with both codes (Eide *et al.*, 2010). In our current investigation, we utilized EPR NLP information to determine sensitivity, specificity, NPVs and PPVs, further advancing the establishment of methodology for ascertaining NMSC in claims data. Differences in PPVs can partly be attributed to difficulty with administrative data date information, which does not always correpond to actual practice. Our findings validate claims data, but also highlight its limitations, suggesting that NLP may be a more accurate ascertainment method.

We are excited to present the validation of NMSC billing claims data against a gold standard–type quality data source. The absence of a population-based tumor registry has hampered the evolution of administrative claims data to study NMSC. As a comprehensive health system, we were able to use NLP and our electronic histopathology reports, which routinely report other ''reportable'' tumors to our tumor registry and the local SEER registry, to identify NMSC, and this EPR information was then utilized as a standard to determine sensitivity and specificity. We believe that this study makes an important additional contribution to establishing validated, accepted methods for ascertaining cases of NMSC with secondary data analysis, as well as highlighting their limitations. Although the implementation of ICD-10 will provide better claims data estimates of SCC and BCC impact, we believe that our study supports using caution when interpreting claims data for NMSC ascertainment: claims data may significantly overestimate actual disease burden, with up to half of ICD-9 ascertained cases found to be false.

We note several limitations to our investigation. This study is limited to a single institution, and should be validated at other institutions or data sets to ensure that it is generalizable. Because we have an open health system, there is the possibility of incomplete claims from patients referred from outside clinicians to the health system or health-plan patients who elected for treatment by an outside, non-HMO provider. Although this is an issue in any open-access US health system, in a subset, we limited our HMO enrollee analysis to patients who had continuous health-plan enrollment during the period of interest to minimize this potential. Historically, NLP can be hampered by negation errors; however, in a setting such as our HMO system, which strives for standardized reporting, our NLP case ascertainment was very robust and validated against a clinic log registry. Finally, the low specificity of the use of administrative claims data using ICD-9-CM, especially in all-payer claims data may be partially due to the possibility of an intervening visit (between biopsy and definitive treatment procedure). This limitation would not be expected to improve with the further implementation of ICD-10 in the United States. These intervening visits would not impact EPR ascertainment and may partly contribute to the superior specificity of EPR NLP (Eide *et al.*, 2010).

Conclusions

We present our findings demonstrating the sensitivity, specificity, and predictive values of administrative claims data algorithms to ascertain NMSC with validation by comparison with the EPRs of a large health system. Considering the substantial difference in cases identified by EPR NLP compared with claims data, we suggest that formal surveillance efforts at the state or national level should be considered and readdressed, as expansion of ICD-9-CM codes in ICD-10 to include unique identifiers for BCC and SCC will not equate to SEER or other tumor registry surveillance accuracy. These algorithms need to be evaluated in other settings and institutions, ideally with similar capacity for validation against electronic histopathology information. The use of EPR NLP in a setting such as the Cancer Research Network's large, diverse population-based, HMO consortium may be a potential alternative to a traditional registry.

MATERIALS AND METHODS

Patients were identified from outpatient health-plan administrative claims data from a large southeastern Michigan HMO and from an outpatient database of individuals with other means of payment seen by the same health-system providers belonging to a salaried medical group. The health system, which consists of 6 hospitals and 32 ambulatory clinics dispersed throughout a tri-county area, reflects the diverse population of the metropolitan area, with the following exceptions. With 13% of health-plan enrollees over the age of 65 years and 30% younger than age 24 years old, the HMO population has a large working-age population, with corresponding modest increases in full-time employment status, household income, and improved general health. As of 2006, the staff model health plan used for this study had an enrollment of 295,000, with a 1-year retention of 84% and a 5-year retention of 56% (National Cancer Institute, 2010). This HMO is a member of the Cancer Research Network (CRN), which is a consortium of integrated health-care systems who have joined efforts for the conduction of collaborative research on preventive, curative, and supportive interventions for major cancers among diverse populations and health systems. The CRN, which was established in 1999, currently consists of 14 health plans, with nearly 11 million enrollees, and is distinguished by their longstanding commitment to prevention and research. Further detail of this HMO, health-plan enrollee demographic information, and generalizability to the surrounding communities has been previously described (Eide *et al.*, 2010). This study was approved by expedited review by the institutional review board.

The gold-standard comparison for algorithm validation was obtained from the EPRs of the health system. Total pathology specimen estimates between 1 January 2007 and 31 December 2008 were obtained leveraging the NLP algorithm within the CoPathPlus Anatomic Pathology Laboratory Information System3 (Sunquest Information Systems, Tucson, AZ). The natural language query combined both structured data fields and free text query, including negation statements and combinatorial algebraic SQL statements. A controlled vocabulary was used and included text strings that matched the diagnostic entities of interest, as well as pertinent tissue types in combination, to create higher specified data returns. The text query was limited to skin tissue samples only for inclusion, and no negation statements were necessary. Each individual query was

combined with subsequent queries using ''or'' statements to ascertain all cases of interest. From CoPathPlus, the following cutaneous malignancies were then identified using free text retrieval capacity from the final diagnosis cell, using the following terms: ''Basal cell carcinoma'' (including ''Fibroepithelioma of Pinkus'' also known as (AKA) ''Pinkus Tumor''), ''Microcystic adnexal carcinoma,'' ''Basosquamous carcinoma,'' ''Squamous cell carcinoma'' (including ''Clear cell squamous cell carcinoma'' AKA ''clear cell carcinoma of the skin'', ''Spindle cell squamous cell carcinoma'' AKA ''spindle cell carcinoma,'' and ''Marjolin's ulcer'' and ''keratoacanthoma''), ''Verrucous carcinoma'' (including ''Carcinoma cuniculatum'' and ''Ackerman tumor''), and Squamous cell carcinoma *in situ* (including ''Bowen disease'', ''Bowen's disease'' and ''Erythroplasia de queyrat''). The text resulting from the CoPath query was further processed using SAS (SAS Institute, Cary, NC) to format data and apply coding logic and classify by histologic type. A sample of EPR data capture was cross-validated against a hardcopy case-log registry book (''clinic log'') maintained at one clinic site within the health system.

Outpatient cases of NMSC for the study period were ascertained from outpatient administrative claims data using ICD-9-CM diagnosis and CPT procedural code algorithms (Eide *et al.*, 2010). The ICD-9-CM diagnosis (for malignant neoplasm of the skin) and CPT procedural code (for excision malignant lesion, destruction of malignant lesion, and chemosurgery/Mohs micrographic technique) algorithms and characteristics of false-positive cases has been previously described in detail; please refer Eide *et al.* (2010) for full description and definition of codes utilized. ICD9 and CPT4 claims data entries were matched to the EHR by visit number to identify incident cases.

Sensitivity, specificity, NPV, and PPV of each algorithm along with corresponding 95% confidence intervals were calculated and compared with the gold standard electronic pathology record. Analyses examined all-payer cases (regardless of health plan), health-plan enrollee's cases only, and the cases ascertained from the clinic log.

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

This study was funded as a pilot award proposal through the National Cancer Institute to the Cancer Research Network (U19 CA 79689). Dr Eide was also supported by a Dermatology Foundation Career Development Award in Health Care Policy.

Author Contributions

All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: MJE, RK, JMT, CCJ. Acquisition of data: MJE, RK, ML, JMT. Analysis and interpretation of data: MJE, JMT, RK, ML, GJ, CCJ. Drafting of the manuscript: MJE. Critical revision of the manuscript for important intellectual content: MJE, JMT, RK, GJ, ML, CCJ. Study supervision: MJE, JMT, CCJ. Statistical analysis: GJ.

REFERENCES

International Agency For Research On Cancer (2008) *Cancer Incidence in Five Continents* IX. IARC Scientific Publications: Lyon, France, pp 1–837

National Cancer Institute (2010) *The HMO Cancer Research Network: Capacity, Collaboration, and Investigation*. National Cancer Institute, April 2010, NIH publication no. 10-6448

Athas WF, Hunt WC, Key CR (2003) Changes in nonmelanoma skin cancer incidence between 1977–1978 and 1998–1999 in Northcentral New Mexico. *Cancer Epidemiol Biomarkers Prev* 12:1105–8

Eide MJ, Krajenta R, Johnson D *et al.* (2010) Identification of patients with nonmelanoma skin cancer using health maintenance organization claims data. *Am J Epidemiol* 171:123–8

Housman TS, Feldman SR, Williford PM *et al.* (2003) Skin cancer is among the most costly of all cancers to treat for the Medicare population. *J Am Acad Dermatol* 48:425–9

Jemal A, Siegel R, Xu J *et al.* (2010) Cancer statistics. *CA Cancer J Clin* 60:277–300

Lamberg AL, Cronin-Fenton D, Olesen AB (2010) Registration in the Danish regional nonmelanoma skin cancer dermatology database: completeness of registration and accuracy of key variables. *Clin Epidemiol* 2:123–36

Rogers HW, Weinstock MA, Harris AR *et al.* (2006) Incidence estimate of nonmelanoma skin cancer in the United States. *Arch Dermatol* 146:283–7