Published in final edited form as:
J Invest Dermatol. 2012 August ; 132(8): 2005–2009. doi:10.1038/jid.2012.98.

Validation of claims data algorithms to identify nonmelanoma skin cancer

Melody J. Eide, MD MPH[1,2], **J. Mark Tuthill, MD MS**[3], **Richard Krajenta, BS**[2], **Gordon Jacobsen, MS**[2], **Marc Levine**[3], and **Christine C Johnson, MPH PhD**[2]

[1]Department of Dermatology, Henry Ford Hospital, Detroit, MI, USA

[2]Department of Public Health Sciences, Henry Ford Hospital, Detroit, MI, USA

[3]Department of Pathology, Henry Ford Hospital, Detroit, MI, USA

Abstract

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 1/1/07 to 12/31/08 were identified from claims data using three ascertainment strategies. These claims-data cases were then compared to NMSC identified using natural language processing (NLP) of electronic pathology reports (EPR), and sensitivity, specificity, positive (PPV) and negative predictive values (NPV) calculated. Comparison of claims data ascertained cases to the NLP demonstrated sensitivities ranging from 48-65% and specificities from 85-98%, with ICD-9-CM ascertainment demonstrating the highest case sensitivity though 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 sensitivity of 98% and specificity of 99%. Validation of administrative data to ascertain NMSC demonstrates respectable sensitivity and specificity though NLP ascertainment was superior. There is a substantial difference in cases identified by NLP compared to claims data suggesting that formal surveillance efforts should be considered.

Address correspondence & Reprint Requests to: Melody J. Eide, MD MPH Henry Ford Hospital Department of Dermatology 3031 West Grand Blvd., Suite 800, Detroit, MI 48202 USA Phone: 313-916-2171; Fax: 313-916-1477; meide1@hfhs.org.

Conflicts of Interest: None

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: Eide MJ, Krajenta R, Tuthill JM, Johnson CC
Acquisition of data: Eide MJ, Krajenta R, Levine M, Tuthill JM
Analysis and interpretation of data: Eide MJ, Tuthill JM, Krajenta R, Levine M, Jacobsen G, Johnson CC
Drafting of the manuscript: Eide MJ
Critical revision of the manuscript for important intellectual content: Eide MJ, Tuthill, JM Krajenta R, Jacobsen G, Levine M, Johnson CC
Study supervision: Eide MJ, Tuthill JM, Johnson CC.
Statistical analysis: Jacobsen G

Keywords

Nonmelanoma Skin Cancer; claims data; surveillance

Introduction

Skin cancer is becoming an increasing health burden.(Athas *et al.*, 2003; Housman *et al.*, 2003a; Rogers *et al.*) The majority of these skin cancers are basal cell (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 (US). Annual incidence of NMSC is estimated to be nearly equal to the incidence of all other cancers combined.(Housman *et al.*; Jemal *et al.*)

Previously we defined and compared algorithms for identifying NMSC using the computerized administrative claims-based dataset 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 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 more than 97% of cases regardless of 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 (EPR). 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 NPV of claims data algorithms to ascertain NMSC cases, with validation against the health system EPR.

Results

From January 1, 2007 to December 31, 2008, there were 24,164 cases involving skin specimens processed by histopathology as identified by the electronic pathology report (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 48-64% and specificities from 85-94%, 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 positive predictive value. (Table 1)

In the HMO population, there were 15,297 total skin specimen cases and 2506 cases of NMSC ascertained from claims data. When compared to the EPR, NMSC claims data algorithms sensitivities ranged from 49-65% and specificities from 96-98%. (Table 2)

One clinic log site, which the EPR was cross validated against, submitted 4614 total cutaneous cases during the study period. This included 909 NMSC cases. The sensitivity, specificity, negative and positive predictive values for the clinic logbook site and the EPR compared favorably, with all values greater than 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 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 to 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 and negative and positive predictive values, 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 non-melanoma skin cancer (NMSC), noting limited registries capturing BCC and SCC, especially in North America.(2008a) 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.*) We believe collaborative efforts of large US HMOs, such as those which 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.(2008b)

We previously reported algorithms for identifying NMSC using the computerized administrative claims-based dataset of this same large US health care system provider and its affiliated HMO.(Eide *et al.*, 2010) NMSC patients who were diagnosed between 1988-2007 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 were reviewed to verify NMSC diagnosis, including all HMO-enrollee members EMRs in 2007, and positive predictive values (PPV) for NMSC were calculated (with sensitivity, specificity and negative predictive value unable to be predicted). A random sample of all years, and all payers were selected for chart review, with PPVs of 47.0% of ICD-9-CM-identified patients, 73.4% of 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 and negative predictive values, further advancing the establishment of methodology for ascertaining NMSC in claims data. Differences in PPVs can in part 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 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. While the implementation of ICD-10 will provide a better claims data estimates of SCC and BCC impact, we believe our study supports utilizing 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 datasets 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. While 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 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 US in 2013. These intervening visits would not impact EPR ascertainment and may in part 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 to the electronic pathology reports of a large health system. Considering the substantial difference in cases identified by EPR NLP compared to 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 in 2013 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 health maintenance organization (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 age 65 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 one-year retention of 84% and a five-year retention of 56%.(2008b) 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 currently consists of 14 health plans, with nearly 11 million enrollees and is distinguished by their longstanding commitment to prevention and research, and was established in 1999. Further detail of this HMO, health plan enrollee demographic information and generalizability to the surrounding communities has been 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 electronic pathology reports (EPR) of the health system. Total pathology specimen estimates between January 1, 2007 and December 31, 2008 were obtained leveraging the natural language processing (NLP) algorithm within the CoPathPlus Anatomic Pathology Laboratory Information System3 (Sunquest Information Systems, Tucson AZ). The natural language query combined both structured data fields as well free text query, including negation statements and combinatorial algebraic SQL statements. A controlled vocabulary was utilized 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, utilizing the following phrases: "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 Inc., Cary, NC, USA.) 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 reference for full description and definition of codes utilized. (Eide *et al.*, 2010) ICD9 and CPT4 claims data entries were matched to the EHR by visit number to identify incident cases.

Sensitivity, specificity, negative (NPV) and positive predictive value (PPV) of each algorithm along with corresponding 95% confidence intervals were calculated, compared to the gold stand 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.

Acknowledgements

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.

Abbreviations

BCC	basal cell carcinoma
CPT	Current Procedural Terminology code
EPR	electronic pathology reports
HMO	health maintenance organization
ICD-9-CM	International Classification of Disease code-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

References

Cancer incidence in five continents. IARC Sci Publ. 2008a; Volume IX:1–837.

The HMO Cancer Research Network: Capacity, Collaboration, and Investigation. National Cancer Institute; 2008b. p. 40

Athas WF, Hunt WC, Key CR. Changes in nonmelanoma skin cancer incidence between 1977-1978 and 1998-1999 in Northcentral New Mexico. Cancer Epidemiol Biomarkers Prev. 2003; 12:1105–1108. [PubMed: 14578151]

Eide MJ, Krajenta R, Johnson D, Long JJ, Jacobsen G, Asgari MM, et al. Identification of Patients With Nonmelanoma Skin Cancer Using Health Maintenance Organization Claims Data. Am J Epidemiol. 2010; 171:123–128. [PubMed: 19969529]

Housman TS, Feldman SR, Williford PM. Skin cancer is among the most costly of all cancers to treat for the Medicare population. J Am Acad Dermatol. 2003a; 48:425–429. [PubMed: 12637924]

Housman TS, Feldman SR, Williford PM, Fleischer AB Jr. Goldman ND, Acostamadiedo JM, et al. Skin cancer is among the most costly of all cancers to treat for the Medicare population. J Am Acad Dermatol. 2003b; 48:425–429. [PubMed: 12637924]

Jemal A, Siegel R, Xu J, Ward E. Cancer statistics. CA Cancer J Clin. 2010; 60:277–300. [PubMed: 20610543]

Lamberg AL, Cronin-Fenton D, Olesen AB. Registration in the Danish Regional Nonmelanoma Skin Cancer Dermatology Database: completeness of registration and accuracy of key variables. Clin Epidemiol. 2:123–136. [PubMed: 20865110]

Rogers HW, Weinstock MA, Harris AR, Hinckley MR, Feldman SR, Fleischer AB, et al. Incidence estimate of nonmelanoma skin cancer in the United States. Arch Dermatol. 2006; 146:283–287. [PubMed: 20231499]

Table 1

Identified and Confirmed Nonmelanoma Skin Cancer Cases with Sensitivity, Specificity, Negative (NPV) and Positive Predictive Values (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	4883	5995	3128	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	4883	5541	3078	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	4883	3441	2335	47.8 (46.4-49.2)	94.3 (93.9-94.6)	87.7 (87.3-88.2)	67.9 (66.3-69.4)

NPV, Negative Predictive Value; PPV, Positive Predictive Value; CI, Confidence Interval; ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; CPT, Current Procedural Terminology.

Study Population: large integrated health system, All-payer Health Plan patients, southeastern MI USA.

Standard: Co-Path3 electronic histopathology data.

Total Skin Pathology Cases: 24, 164

Table 2

Identified and Confirmed Nonmelanoma Skin Cancer Cases with Sensitivity, Specificity, Negative (NPV) and Positive Predictive Values (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	2506	2209	1639	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	2506	2059	1610	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	2506	1534	1230	49.1 (47.1-51.0)	97.6 (97.4-97.9)	90.7 (90.2-91.2)	80.2 (78.2-82.2)

HMO, Health Maintenance Organization; PPV, Positive Predictive Value; CI, Confidence Interval; ICD- 9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; CPT, Current Procedural Terminology.

Study Population: large integrated health system, HMO Health Plan Enrollees, southeastern MI USA.

Standard: Co-Path3 electronic histopathology data.

Total Skin Pathology Cases: 15,297

Table 3

Confirmed and Identified Nonmelanoma Skin Cancer Cases with Sensitivity, Specificity, Negative (NPV) and Positive Predictive Values (PPV), 2007-2008, Electronic Pathology record and Log- book 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)

TFNEPR, electronic histopathology record; NPV, Negative Predictive Value; PPV, Positive Predictive Value; CI, Confidence Interval.

Study Population: large integrated health system, All-payer Health Plan patients, logbook clinic site, southeastern MI USA with Co-Path3 electronic histopathology record (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, Negative (NPV) and Positive Predictive Values (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)

NPV, Negative Predictive Value; PPV, Positive Predictive Value; CI, Confidence Interval; ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; CPT, Current Procedural Terminology.

Study Population: large integrated health system, single clinical site, southeastern MI USA.

Standard: Co-Path3 electronic histopathology data.

Total Skin Pathology Cases: 4, 614

Table 5

Characteristics of False Cases of Nonmelanoma Skin Cancer (NMSC) Identified by either Clinic Log or Electronic HistoPathology (EHP) information, 2007-2008.

Reason	**Unconfirmed cases**	
	EPR (N=15)	**Clinic Log (N=16)**
Missing information (Omission or abstractor error possible)	5	0
Basal cell or Squamous cell carcinoma*	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,”etc.)	1	2
Exclusion of cutaneous malignancy (e.g. “Scar, negative for”, “Negative for”, “No residual,” etc.)	6	0
Different physical clinical location that clinic log site itself *	1	1

NMSC, Nonmelanoma skin cancer; EPR, electronic histopathology record.

Study Population: large integrated health system, single clinical site and Co-Path3 electronic histopathology record data, southeastern MI USA.

*Patient was flagged as present in both columns due to nonmatching date of service differences (N=2; of which 1 had different clinic site and 1 was a squamous cell carcinoma)