Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2013 November ; 22(11): . doi:10.1002/pds.3520.

Validation of claims-based algorithms for identification of high-grade cervical dysplasia and cervical cancer

Seouyoung C Kim, MD, MSCE^{1,2}, Victoria G Gillet, BA, Sarah Feldman, MD, MPH³, Huichuan Lii, MA, MS, Sengwee Toh, ScD⁴, Jeffrey S Brown, PhD⁴, Jeffrey N Katz, MD, MS², Daniel H Solomon, MD, MPH^{1,2}, and Sebastian Schneeweiss, MD, SD¹

¹Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital/Harvard Medical School, Boston, MA

²Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA

³Gynecologic Oncology, Dana-Farber Cancer Institute, Boston, MA

⁴Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA

⁵Department of Orthopedic Surgery, Brigham and Women's Hospital/Harvard Medical School, Boston, MA

Abstract

Background—High-grade cervical dysplasia or cervical intraepithelial neoplasia (CIN) grade 2 or worse has been widely used as a surrogate endpoint in cervical cancer screening or prevention trials.

Methods—To identify high-grade cervical dysplasia and cervical cancer, we developed claims-based algorithms that incorporated a combination of diagnosis and procedure codes using the billing data in an electronic medical records (EMR) database and assessed the validity of the algorithms in an independent administrative claims database. We calculated the positive predictive value (PPV) with the 95% confidence interval (CI) of each algorithm, using new cytologic or pathologic diagnosis of CIN 2 or 3, carcinoma in situ, or cervical cancer as the gold standard.

Results—Having 1 diagnosis code for high-grade cervical dysplasia or cervical cancer had a PPV of 57.1% (95%CI 54.7–59.5%). By requiring 2 diagnoses for high-grade cervical dysplasia or cervical cancer, separated by 7 to 30 days, the PPV increased to 60.2% (95%CI 53.9–66.1%). At least 2 diagnoses and a procedure code within a month from the first diagnosis date yielded a PPV of 80.7% (95%CI 73.6–86.2%). The algorithms had greater PPVs in identifying prevalent high-grade cervical dysplasia or cervical cancer. Overall, the PPVs of these algorithms were similar or slightly lower in the external claims data than in the sample used to derive the algorithms.

Conclusions—Use of 2 diagnosis codes in combination with a procedure code appears to be a valid tool for studying high-grade cervical dysplasia and cervical cancer in both EMR and administrative claims databases.

Corresponding Author: Seouyoung C. Kim, MD, MSCE, Address: 1620 Tremont Street, Suite 3030, Boston MA 02120, USA, Phone: 1-617-525-8863, Fax: 1-617-232-8602, skim62@partners.org.

Reprint Request to: Seouyoung C Kim, MD, MSCE, Address: 1620 Tremont Street, Suite 3030, Boston MA 02120, USA

Disclosures:

Gillet, Feldman, Lii and Toh have nothing to disclose.

Keywords

cervical dysplasia; cervical cancer; validity; claims data

INTRODUCTION

More than 100 different types of human papillomavirus (HPV) have been identified and high-risk types (e.g., HPV-16 and 18) can cause precancerous cervical dysplasia and potentially fatal, invasive cervical cancer.[1, 2] Cervical cancer screening with a Papanicolaou (Pap) test has been widely accepted and has reduced the incidence and prevalence of cervical cancer. While cervical cancer is still the third most common cancer in women worldwide, [3] it is quite rare in the U.S. with the annual incidence of 8 per 100,000 women.[4]

In the past, the terms, mild, moderate, and severe dysplasia, were used to describe the severity of cervical dysplasia seen in Pap tests. After several modifications in the nomenclature to classify cytological and histological cervical changes associated with HPV and precancerous lesions, the current classification system uses cervical intraepithelial neoplasia (CIN) for histological changes and cervical squamous intraepithelial lesions (SIL) for cytological changes.[5] High grade lesions include CIN 2, formerly moderate dysplasia, which refers to dysplastic cellular changes confined to the basal two-thirds of the epithelium and CIN 3, formerly severe dysplasia and carcinoma in situ, which refers to dysplastic cellular changes encompassing greater than two-thirds of the epithelial thickness, and includes full-thickness lesions. The mean reporting rate of high grade SIL in U.S. laboratories is about 1%. [6] Most patients with high grade lesions undergo interventions such as ablative or excisional procedures for further diagnosis and treatment.[7, 8]

Large-scale health care utilization databases have been widely used in epidemiologic and pharmacoepidemiologic research and can be particularly useful in conducting a population-based study of rare outcomes such as cervical cancer.[9] Many cervical cancer screening or prevention trials have used high grade cervical dysplasia (CIN 2 or worse) as a surrogate endpoint in order to decrease sample size and follow-up duration, ultimately reducing resource requirements.[10–12] However, the role of electronic health care utilization databases for research on such surrogate endpoint, high-grade cervical dysplasia, related to use of vaccines or drugs safety has been limited and the extent of the potential misclassification bias in defining high-grade cervical dysplasia based on the diagnosis and/or procedure codes has not been studied. Therefore, we sought 1) to develop claims-based algorithms to identify high-grade cervical dysplasia and cervical cancer in the electronic medical records (EMR) of a large academic facility and 2) to conduct a validation study in an external administrative claims database linked to the EMR of a large, multispecialty medical practice.

METHODS

Data Sources

For algorithm development, we used the billing data from the Partners' Research Patient Data Registry (RPDR).[13] The RPDR is a centralized clinical data warehouse and contains data from Partners' EMR as well as billing information from inpatient and outpatient visits. For the external validation of the algorithm, we used the Harvard Pilgrim Health Care's (HPHC) administrative claims database linked to the EMR of Harvard Vanguard Medical Associates (HVMA), a large, multispecialty medical group practice serving Eastern Massachusetts. HPHC is a non-profit health plan serving the New England area and one of

the 18 health care delivery organizations for the HMO Research Network.[14, 15] This work was approved by the Institutional Review Board at Brigham and Women's Hospital (BWH) and the Human Studies Committee at the HPHC.

Algorithm Development

We developed several algorithms to identify patients with diagnoses of high-grade cervical dysplasia or cervical cancer using a combination of diagnosis and procedure codes in the billing data of the RPDR. Procedure codes for various types of procedures for further diagnosis or management of high-grade cervical dysplasia were included in the algorithms. [8] Table 1 shows the list of diagnosis and procedure codes that we used in the algorithms. In the RPDR (2007–2009), female patients aged between 20 and 60 years who had either an outpatient claim coded for any of abnormal Pap smear tests (ICD-9 code: 795.0x, 622.1x, 616.9, 233.x, 180.x, and V73.81) were selected as a study cohort for development of the algorithms. Cytologic or pathologic results of high-grade cervical dysplasia (i.e., high grade SIL, CIN 2 or worse) were used as the gold standard to validate the claims-based algorithms. The confirmed high-grade cervical dysplasia within 30 days of the algorithm-identified diagnosis date in the billing data were considered as “incident case”. Confirmed diagnoses of high-grade cervical dysplasia any time prior to the algorithm-identified diagnosis date were considered as “prevalent case”.

The positive predictive value (PPVs) was calculated as the percentage of the patients who met the pathologic definitions of high-grade dysplasia or cervical cancer among those identified by the respective algorithms. Ninety-five per cent confidence intervals (CIs) of the PPV for each algorithm were calculated by using the normal approximation of the binomial distribution.

Algorithm Validation

Several algorithms developed in the RPDR were tested in the HPHC claims database (2008–2010), to identify female patients aged 20 to 60 years old with high-grade cervical dysplasia or cervical cancer. Patients were required to have a continuous enrollment 3 months prior to and after the algorithm-identified diagnosis date. The algorithm-identified cases were confirmed based on the cytologic or pathologic results from the HVMA's EMR linked to the claims database. Only the incident cases, diagnosed within 30 days from the algorithm-identified diagnosis date, were validated.

We conducted a hypothetical sensitivity analysis to illustrate the potential impact of outcome misclassification. A hypothetical cohort was constructed to study the effect of a specific exposure on the risk of high-grade cervical dysplasia and cervical cancer. Let us suppose there are 200 patients with high-grade cervical dysplasia or cervical cancer in the total study population of 20,000 (i.e., prevalence of 1% [6]) and the true relative risk (RR) and risk difference (RD) of high-grade cervical dysplasia or cervical cancer in patients exposed to a certain disease or drug is 2.33 and 0.008 (Table 4a), respectively. We then estimated RR and RD for two separate algorithms based on a combination of different sensitivity and specificity.

All analyses were performed using SAS 9.3 Statistical Software (SAS Institute Inc., Cary, NC).

RESULTS

There were a total of 24,426 female patients who had at least one abnormal Pap smear between 2007 and 2009 in the RPDR. Of those, 1,713 (7%) patients were identified by the algorithms as having at least one diagnosis of high-grade cervical dysplasia or cervical

cancer. The mean age (SD) was 32.0 (9.4) years. 57.1% of these patients had a confirmed incident cases of CIN 2 or worse within 30 days of the algorithm-identified diagnosis date and 67.5% had a prevalent case of CIN 2 or worse. The algorithm that required at least two diagnoses and one procedure code identified 150 patients (0.6%) with the mean age (SD) of 32.2 (8.6) years. The PPV of this algorithm was 80.7% for an incident case and 90.7% for a prevalent case of CIN 2 or worse. Table 2 summarizes the number of patients identified by each algorithm, the number of confirmed incident and prevalent cases, and the PPVs with 95% CIs. The algorithm requiring at least two diagnoses and one procedure code identified 150 women. Of those, 136 patients had a prevalent case which is a confirmed diagnosis of high-grade cervical dysplasia or cervical cancer any time prior to the algorithm-identified diagnosis date. 121 patients had an incident case defined as a confirmed diagnosis of high-grade cervical dysplasia or cervical cancer within 30 days of the algorithm-identified diagnosis date.

In the HPHC claims data, the PPVs of these algorithms were similar or slightly lower (Table 3). There were 148 female HVMA patients with at least one diagnosis for high-grade cervical dysplasia or cervical cancer and at least one procedure code; 96 (PPV=68.6%) of these patients had a confirmed case based on review of their medical record and their mean age (SD) was 32.6 (9.1) years. The algorithm that required at least two diagnoses and one procedure code identified 106 female patients; of those, 75 had the confirmed high-grade cervical dysplasia or cervical cancer (PPV=75.0%) based on review of their medical record and their mean age (SD) was 33.6 (9.7) years. Further restriction on the duration between the two diagnosis dates to be 7 to 30 days identified 43 patients and PPV was 74.4%.

In our hypothetical sensitivity analysis, substantial bias towards the null was observed on estimates of both RR and RD, due to non-differential misclassification of the outcome in these data. For example, when the specificity of the algorithm requiring two diagnoses from two separate visits which has a PPV of 60.2% is assumed to be 99.93%, the estimated sensitivity is 10%. With these test characteristics, only 34 patients ($200 \times \text{sensitivity} + 19,800 \times [1 - \text{specificity}]$) would be identified as having high-grade cervical dysplasia or cervical cancer.[16] The more restrictive algorithm requiring two diagnoses for high-grade cervical dysplasia or cervical cancer and one procedure code which has a PPV of 80.7% would yield the RR of 1.95 and the RD of 0.0008, assuming the sensitivity of 10% and the specificity of 100%.

DISCUSSION

We developed claims-based algorithms to identify high-grade cervical dysplasia and cervical cancer using the EMR billing data and validated it in an independent claims database. Our study demonstrated that the algorithm combining at least two diagnosis codes for high-grade cervical dysplasia or cervical cancer and at least 1 procedure code related to further diagnosis or treatment of cervical dysplasia could be a valid tool in an epidemiologic study using electronic health care databases.

To our knowledge, no prior study has developed or validated a claims-based algorithm for high-grade cervical dysplasia. Unlike other invasive cancers, most women get diagnosed with cervical dysplasia and treated before they fully develop invasive cervical cancer because of the broad use of routine Pap smear screening. Population-level data on high grade cervical dysplasia is not available in most cancer registries such as the Surveillance Epidemiology and End Results. With our claims-based algorithms, large-scale population-based data sources such as insurance claims databases can be considered for epidemiologic or pharmacoepidemiologic studies on the risk of high-grade cervical dysplasia and cervical cancer.

Our study has a few limitations. First, the sensitivity and specificity of the algorithms were not assessed in this study. As the PPV is related to the prevalence of the disease, it could vary from one population to another. However, we used two different patient populations with potentially different prevalence of high grade cervical dysplasia, a tertiary academic center and a large community-based multispecialty practice, and observed similar PPVs for the algorithms. Therefore, our findings are likely to be generalizable to other databases. Second, due to the rarity of high grade cervical dysplasia, the performance of the algorithms on identifying patients with persistent high-grade cervical dysplasia was not assessed. Third, although cytologic high grade squamous intraepithelial lesion (HSIL) is not necessarily equivalent to histologic CIN 2 or 3, we considered both HSIL or CIN 2 or 3 the same.

In summary, despite the substantial bias toward the null due to non-differential outcome misclassification, the use of at least two diagnosis codes in combination with a procedure code for high-grade cervical dysplasia and cervical cancer appears to be a valid tool for studying clinical and epidemiological issues related to this uncommon outcome, high-grade cervical dysplasia and cervical cancer, in electronic health care databases.

Acknowledgments

Kim is supported by the NIH grant K23 AR059677. She received research support from Pfizer and tuition support for the Pharmacoepidemiology Program at the Harvard School of Public Health funded by Pfizer and Asisa.

Brown receives funding from several federal agencies, including the HMO Research Network Center for Education and Research on Therapeutics (AHRQ, 1U18HS016955-1) and the HMO Research Network DECIDE Center, both funded by the Agency for Healthcare Research and Quality. Brown receives consulting fees from Oracle Corp. and Scimetrika LLC.

Katz receives support from several NIH/NIAMS grants including P60 AR 047782 as well as the Informed Medical Decisions Foundation.

Solomon is supported by the NIH grants K24 AR055989, P60 AR047782, and R01 AR056215. Solomon has received research support from Amgen, Lilly, and CORRONA. He serves in unpaid roles on Pfizer and Novartis sponsored trials.

Schneeweiss is principal investigator of the Brigham and Women's Hospital DECIDE Center on Comparative Effectiveness Research, funded by the Agency for Healthcare Research and Quality, and of the Harvard-Brigham Drug Safety and Risk Management Research Contract, funded by the US Food and Drug Administration. Schneeweiss has received consulting fees from WHISCON, LLC and Booz & Company and is principal investigator of research grants from Pfizer, Novartis and Boehringer Ingelheim to the Brigham and Women's Hospital.

REFERENCES

1. Muñoz N, Bosch F, de Sanjosé S, et al. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med*. 2003; 348:518–527. [PubMed: 12571259]
2. Walboomers JM, Jacobs MV, Manos MM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol*. 1999; 189:12–19. [PubMed: 10451482]
3. Ferlay J, Shin H, Bray F, et al. Cervical Cancer Incidence and Mortality Worldwide in 2008. *Cancer Fact Sheet*. 2008 GLOBOCAN 2008 v2.0:[Available from: <http://globocan.iarc.fr/factsheets/cancers/cervix.asp>.
4. Howlader, N.; Noone, AM.; Krapcho, M., et al. SEER Cancer Statistics Review, 1975–2009. Bethesda, MD: National Cancer Institute; 2012. SEER Stat Fact Sheets: Cervix Uteri.
5. Solomon D, Davey D, Kurman R, et al. The 2001 Bethesda System: terminology for reporting results of cervical cytology. *JAMA*. 2002; 287:2114–2119. [PubMed: 11966386]
6. Davey DD, Neal MH, Wilbur DC, et al. Bethesda 2001 implementation and reporting rates: 2003 practices of participants in the College of American Pathologists Interlaboratory Comparison

- Program in Cervicovaginal Cytology. Arch Pathol Lab Med. 2004; 128:1224–1229. [PubMed: 15504056]
7. Mitchell MF, Tortolero-Luna G, Wright T, et al. Cervical human papillomavirus infection and intraepithelial neoplasia: a review. J Natl Cancer Inst Monogr. 1996; 21
 8. Wright TCJ, Massad LS, Dunton CJ, et al. 2006 consensus guidelines for the management of women with abnormal cervical cancer screening tests. Am J Obstet Gynecol. 2007; 197:346–355. [PubMed: 17904957]
 9. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58:323–337. [PubMed: 15862718]
 10. FUTURE II Study Group. Quadrivalent vaccine against human papillomavirus to prevent high-grade cervical lesions. N Engl J Med. 2007 May 10; 356(19):1915–1927. 2007; **356**:1915–27. [PubMed: 17494925]
 11. Mao C, Koutsky L, Ault K, et al. Efficacy of human papillomavirus-16 vaccine to prevent cervical intraepithelial neoplasia: a randomized controlled trial. Obstet Gynecol. 2006; 107:18–27. [PubMed: 16394035]
 12. Schatzkin A, Gail M. The promise and peril of surrogate end points in cancer research. Nat Rev Cancer. 2002; 2:19–27. [PubMed: 11902582]
 13. Partners HealthCare System. Research Patient Data Registry (RPDR). 2012 Available from: <http://rc.partners.org/rpdr>.
 14. HMO Research Network. Members. Available from: <http://www.hmoresearchnetwork.org/members.htm>.
 15. Moore KM, Duddy A, Lee GM, et al. Outpatient urticaria diagnosis codes have limited predictive value for same-day influenza vaccine adverse event detection. J Clin Epidemiol. 2010 Apr. 63:407–411. [PubMed: 19889513]
 16. Rothman, KJ.; Greenland, S.; Lash, TL. Validity in epidemiologic studies. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. Modern Epidemiology. 3rd ed.. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 137–145.

Key Points

- High-grade cervical dysplasia has been widely used as a surrogate endpoint for cervical cancer.
- Use of at least two diagnosis codes in combination with a procedure code appears to be a valid tool for studying high-grade cervical dysplasia in electronic health care database.

Table 1

Cervical dysplasia related-diagnosis and procedure codes

ICD-9 code	Description	CPT code	Description
795.04	High-grade squamous intraepithelial lesion	57420–57421	Colposcopy of the entire vagina, with and without biopsy
622.12	Moderate dysplasia or CIN II	57452	Colposcopy of the cervix
233.1	Carcinoma in situ of cervix or CIN III	57454	Colposcopy of the cervix, with biopsy of the cervix and ECC
180.1	Malignant neoplasm of cervix	57455	Colposcopy of the cervix, with biopsy
		57456	Colposcopy of the cervix, with ECC
		57460	Colposcopy of the cervix, with loop electrode biopsy
		57461	Colposcopy of the cervix, with loop electrode conization
		57500	Cervical biopsy
		57505	ECC
		57511	Cryotherapy of Cervix
		57510	Electro or thermal cautery of cervix:
		57513	Laser ablation
		57520	Cervical conization
		57522	Loop excision
		57530–57531	Cervicectomy
		57540–57550	Excision of cervical stump

ICD-9: International Classification of Diseases, 9th Revision, CPT: Current Procedural Terminology, CIN: cervical intraepithelial neoplasia, ECC: endocervical curettage

Positive predictive value (PPV) and 95% confidence interval of the algorithms: Algorithm development in the Partners' Research Patient Data Registry database

Table 2

Algorithm	Records identified	Pathology not available	Incident case		Prevalent case	
	N	N	N	PPV % (95% CI)	N	PPV % (95% CI)
1 diagnosis for cervical cancer or high-grade cervical dysplasia ^a	1,713	44	953	57.1 (54.7–59.5)	1127	67.5 (65.3–69.7)
2 diagnoses for cervical cancer or high-grade cervical dysplasia from 2 separate encounters ^b	246	0	148	60.2 (53.9–66.1)	170	69.1 (63.1–74.6)
1 diagnosis for cervical cancer or high-grade cervical dysplasia AND 1 CPT code	809	12	470	59.0 (55.5–62.3)	574	72.0 (68.8–75.0)
2 diagnoses for cervical cancer or high-grade cervical dysplasia AND 1 CPT code	150	0	121	80.7 (73.6–86.2)	136	90.7 (85.0–94.4)

^aThis algorithm was not tested further in the external data source.

^bEncounters are separated by 7–30 days. CPT: Current Procedural Terminology Incident case was defined as a confirmed diagnosis of high-grade cervical dysplasia within 30 days of the algorithm-identified diagnosis date in the billing data. Prevalent case was defined as a confirmed diagnosis of high-grade cervical dysplasia any time prior to the algorithm-identified diagnosis date.

Table 3

Positive predictive value (PPV) and 95% confidence interval of the algorithms: Validation in the Harvard Pilgrim Health Care/Harvard Vanguard Medical Associates database

Algorithm	Records identified	Pathology not available	Incident case	
	N	N	N	PPV % (95% CI)
1 diagnosis for cervical cancer or high-grade cervical dysplasia AND 1 CPT code	148	8	96	68.6 (60.5–75.7)
2 diagnoses for cervical cancer or high-grade cervical dysplasia from 2 separate encounters*	47	4	30	69.8 (54.9–81.4)
2 diagnoses for cervical cancer or high-grade cervical dysplasia AND 1 CPT code	106	6	75	75.0 (65.7–82.5)
2 diagnoses for cervical cancer or high-grade cervical dysplasia AND 1 CPT code*	43	4	29	74.4 (58.9–85.4)

* Encounters are separated by 7–30 days. CPT: Current Procedural Terminology

The impact of outcome misclassification in a hypothetical cohort study assessing the effect of a specific exposure on the risk of high-grade cervical dysplasia and cervical cancer using claims-based algorithms

Table 4

a. 2 by 2 table for a hypothetical data				
	True Cases	Total	True Risk	True RR
Exposed	140	10,000	0.014	2.33
Unexposed	60	10,000	0.006	
Total	200	20,000		

b. The impact of outcome misclassification				
Algorithm	Sensitivity (%)	Specificity (%)	Estimated RR*	Estimated RD*
2 diagnoses for cervical cancer or high-grade cervical dysplasia from 2 separate visits (PPV= 60.2%)	10.00	99.93	1.61	0.0008
	50.00	99.67	1.63	0.004
	80.00	99.47	1.63	0.006
2 diagnoses for cervical cancer or high-grade cervical dysplasia AND 1 CPT code (PPV = 80.7%)	10.00	99.98	1.95	0.0008
	50.00	99.88	1.95	0.004
	80.00	99.81	1.95	0.006

Prevalence of high-grade cervical dysplasia or cervical cancer in the study population was assumed to be 1%.[6]
The sensitivity was assumed. The specificity was estimated based on the observed PPV and the assumed prevalence and sensitivity.
RR= relative risk; RD= risk difference; ICD-9: International Classification of Diseases, 9th revision
* RR and RD were estimated using a claims-based algorithm for outcome identification.