






Improving Cervical Precancer Surveillance: Validity of Claims-Based Prediction Models in ICD-9 and ICD-10 Eras

Jaimie Z. Shing , MPH,¹ Marie R. Griffin, MD, MPH,² Linh D. Nguyen , BS,³ James C. Slaughter , DrPH,⁴ Edward F. Mitchel , MS,² Manideepthi Pemmaraju, MBBS, MPH,² Alyssa B. Rentuza, BS,¹ Pamela C. Hull , PhD, MA^{1,5,*}

¹Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA, ²Department of Health Policy, Vanderbilt University Medical Center, Nashville, TN, USA, ³Data Science Institute, Vanderbilt University, Nashville, TN, USA, ⁴Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA and ⁵Department of Behavioral Science, University of Kentucky Markey Cancer Center, Lexington, KY, USA

*Correspondence to: Pamela C. Hull, PhD, MA, 2365 Harrodsburg Rd, Suite A230, Lexington, KY, USA (e-mail: pam.hull@uky.edu).

Abstract

Background: Human papillomavirus vaccine (HPV) impact on cervical precancer (cervical intraepithelial neoplasia grades 2+ [CIN2+]) is observable sooner than impact on cancer. Biopsy-confirmed CIN2+ is not included in most US cancer registries. Billing codes could provide surrogate metrics; however, the International Classification of Diseases, ninth (ICD-9) to tenth (ICD-10) transition disrupts trends. We built, validated, and compared claims-based models to identify CIN2+ events in both ICD eras. **Methods:** A database of Davidson County (Nashville), Tennessee, pathology-confirmed CIN2+ from the HPV Vaccine Impact Monitoring Project (HPV-IMPACT) provided gold standard events. Using Tennessee Medicaid 2008-2017, cervical diagnostic procedures (N = 8549) among Davidson County women aged 18-39 years were randomly split into 60% training and 40% testing sets. Relevant diagnosis, procedure, and screening codes were used to build models from CIN2+ tissue diagnosis codes alone, least absolute shrinkage and selection operator (LASSO), and random forest. Model-classified index events were counted to estimate incident events. **Results:** HPV-IMPACT identified 983 incident CIN2+ events. Models identified 1007 (LASSO), 1245 (CIN2+ tissue diagnosis codes alone), and 957 (random forest) incident events. LASSO performed well in ICD-9 and ICD-10 eras: 77.3% (95% confidence interval [CI] = 72.5% to 81.5%) vs 81.1% (95% CI = 71.5% to 88.6%) sensitivity, 93.0% (95% CI = 91.9% to 94.0%) vs 90.2% (95% CI = 87.2% to 92.7%) specificity, 61.3% (95% CI = 56.6% to 65.8%) vs 60.3% (95% CI = 51.0% to 69.1%) positive predictive value, 96.6% (95% CI = 95.8% to 97.3%) vs 96.3% (95% CI = 94.1% to 97.8%) negative predictive value, 91.0% (95% CI = 89.9% to 92.1%) vs 88.8% (95% CI = 85.9% to 91.2%) accuracy, and 85.1% (95% CI = 82.9% to 87.4%) vs 85.6% (95% CI = 81.4% to 89.9%) C-indices, respectively; performance did not statistically significantly differ between eras (95% confidence intervals all overlapped). **Conclusions:** Results confirmed model utility with good performance across both ICD eras for CIN2+ surveillance. Validated claims-based models may be used in future CIN2+ trend analyses to estimate HPV vaccine impact where population-based biopsies are unavailable.

The human papillomavirus (HPV) vaccine's impact on cervical precancers, including cervical intraepithelial neoplasia (CIN) grades 2 and 3 and adenocarcinoma in situ (together referred to as CIN2+), may be observed sooner than the vaccine's impact on cervical cancer (1). The HPV vaccine can prevent nearly 80% of CIN2+ (2), and preventing precancers will ultimately prevent cervical cancer and associated premature mortality (3). Additionally, precancers are associated with considerable preventable morbidity and costs (4-6). In the United States, 196 000 CIN2+ events were diagnosed in 2016

(2). Despite declines from 216 000 CIN2+ events in 2008 (2), the HPV vaccine is not yet reaching its full potential. HPV vaccination coverage lags behind the two other recommended adolescent vaccines, which are the meningococcal conjugate vaccine and the combined tetanus, diphtheria, and pertussis vaccine, and there is substantial variation in coverage rates across states (7). Monitoring trends in CIN2+ is critical for evaluating the impact of HPV vaccination over time and targeting vaccine promotion and cervical cancer screening efforts.

Received: 22 July 2020; Revised: 16 October 2020; Accepted: 1 November 2020

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Examining CIN2+ in the United States is challenging. CIN2+ diagnosis confirmation requires cervical biopsies, which are not included in most US cancer registries or surveillance systems. Several states have monitored CIN2+ rates through the state-based Pap registry in New Mexico and the population-based HPV Vaccine Impact Monitoring Project (HPV-IMPACT) in 5 states (2,8-13). However, the vast majority of states do not have such surveillance capacity, so it is not possible to examine national CIN2+ trends or variation across states.

A potential solution is leveraging administrative data using International Classification of Diseases, Clinical Modification (ICD), Current Procedural Terminology, and Healthcare Common Procedure Coding System codes, which systematically classify diseases and patient procedures, providing surrogate metrics. However, the transition from the ninth ICD (ICD-9) to the tenth ICD (ICD-10) coding revision in 2015 (14) disrupts trends because ICD-9 and ICD-10 codes differ in structure (14,15). Compared with ICD-9, ICD-10 codes have more detail about laterality, severity, and complexity of health conditions, allowing for increased specificity and accuracy (15).

Despite the need to expand options for the surveillance of CIN2+ incidence, limited information is available on the validity of claims data for identifying incident CIN2+ events between ICD-9 and ICD-10 eras. To our knowledge, no studies have validated claims-based CIN2+ models that can detect trends in CIN2+ across both ICD eras. Although such models are not intended to provide the highest accuracy that would be needed for clinical decision making, they would be useful for detecting trends in public health surveillance. To address this gap, we aimed to build and validate claims-based models identifying CIN2+ events in ICD-9 and ICD-10 eras as a method to estimate the number of CIN2+ events in the population, and we compared 3 model-building approaches to identify an optimal model. In addition, to provide insight into unifying period continuity across ICD-9 and ICD-10 eras for future trend analyses of HPV vaccine impact, we compared model performance between the 2 ICD eras.

Methods

Study Population

Billing codes from the Tennessee Medicaid program (TennCare) identified women with cervical diagnostic procedural encounters from 2008 to 2017 who were enrolled in TennCare at the time of the procedure (Supplementary Table 1, available online). We included women aged 18-39 years residing in Davidson County (Nashville), Tennessee, because our gold standard dataset for validation had the same age and geographic inclusion (Figure 1). We counted encounters rather than women to account for women with multiple encounters. Procedures within 30 days of each other were considered clusters of associated procedures and counted as 1 encounter; 1453 clusters were identified among 10002 total procedures in 2008-2017 (final sample = 8549 encounters). This research was approved by the Division of TennCare and deemed public health surveillance, which was thereby exempt by the Tennessee Department of Health and Vanderbilt University Institutional Review Boards.

Gold Standard

Biopsy-confirmed CIN2+ events, including CIN2, CIN3, and adenocarcinoma in situ, in Davidson County, Tennessee, were

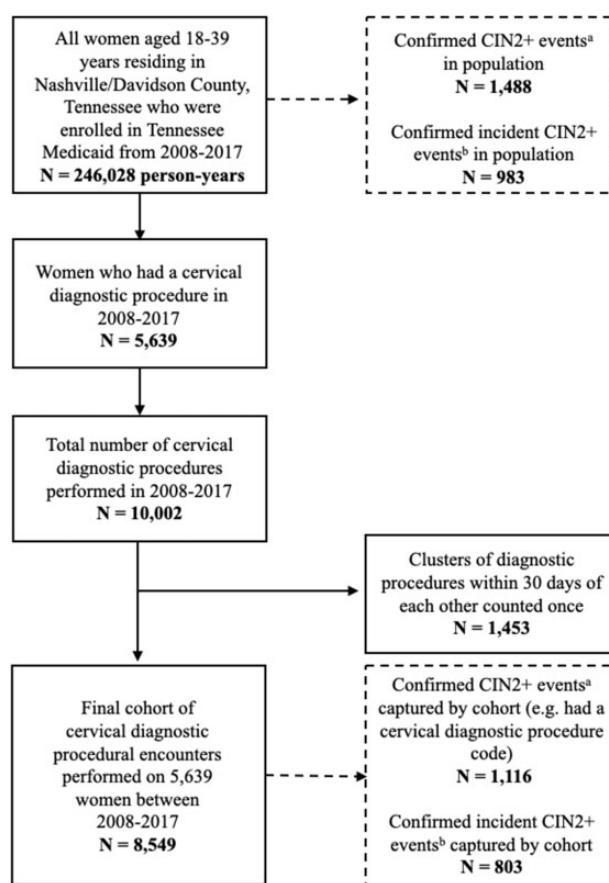


Figure 1. Flow diagram to capture cohort of cervical diagnostic procedural encounters from 2008 to 2017 among Tennessee Medicaid (TennCare)-enrolled women aged 18-39 years residing in Davidson County, Tennessee. ^aConfirmed cervical intraepithelial neoplasia (CIN) grades 2+ (CIN2+) events are all events reported and validated by the Human Papillomavirus Vaccine Impact Monitoring Project (HPV-IMPACT); multiple events may be included for each woman. ^bConfirmed incident CIN2+ events are index events for each woman.

collected and validated by the HPV-IMPACT team at Vanderbilt University Medical Center as part of the HPV-IMPACT monitoring project (16), a program with partnerships between the Centers for Disease Control and Prevention, academic institutions, and Emerging Infections Programs in 5 state health departments (16). HPV-IMPACT conducts enhanced CIN2+ surveillance among women aged 18-39 years in catchment areas, including Davidson County. The HPV-IMPACT team receives reports from pathology laboratories and reviews charts of women with pathologically confirmed CIN2+ to assure these women were Davidson County residents at the time of biopsy and the biopsy reflected an incident event. Records of women with cervical biopsies identified through administrative databases are audited to capture all events. From 2008 to 2017, HPV-IMPACT identified 1488 CIN2+ events among TennCare-enrolled women aged 18-39 years residing in Davidson County, of which 983 were incident events (Figure 1).

In our analytic sample, encounters were considered confirmed events if the diagnostic procedure was from a woman with an HPV-IMPACT confirmed CIN2+ event. We found the interval between these women's diagnostic procedure dates and their closest HPV-IMPACT confirmed diagnosis date ranged from 0 to 3131 days (median = 28 days). Therefore, we used predetermined conservative parameters to associate diagnoses

with their most probable corresponding procedures. Encounters were considered confirmed events only if the confirmed diagnosis date was within ± 60 days of procedure date or within 60 days before the first diagnostic procedure in a cluster (procedures within 30 days of each other) and 60 days after the last diagnostic procedure in the cluster. Given the inclusion criteria and prespecified parameters, 1116 confirmed events were captured in our final sample of cervical diagnostic procedures among TennCare-enrolled women aged 18-39 years residing in Davidson County, of which 803 were incident events (Figure 1).

Predictors

From each diagnostic procedure date, we used the same interval parameters for determining confirmed event status to search for presence of ICD-9, ICD-10, Current Procedural Terminology, and Healthcare Common Procedure Coding System codes relating to a CIN2+ tissue diagnosis, nonspecific CIN tissue diagnosis, high-grade squamous intraepithelial lesion (HGSIL) cytology diagnosis, CIN1 tissue diagnosis, low-grade squamous intraepithelial lesion cytology diagnosis, atypical squamous cells of undetermined significance (ASCUS) diagnosis, HPV screening test, Papanicolaou (Pap) smear or test, HPV DNA test, cervical treatment procedure, and cervical or vaginal biopsy. We consulted content experts to determine appropriate predictor groupings using a single code or combination of codes (Supplementary Table 1, available online).

Model Building

The data were randomly split into 60% training and 40% testing sets, by era. To assess ICD-10 coding implementation lag for the transition cutoff (October 1, 2015), we examined crossover usage of ICD-9 codes after September 30, 2015, and ICD-10 codes before October 1, 2015. Only 16 of 1444 (1.11%) encounters used an ICD-9 code after September 30, 2015, and 5 of 7105 (0.07%) encounters used an ICD-10 code before October 1, 2015. Because crossover was minimal, we retained the original cutoff; ICD-9 and ICD-10 eras consisted of encounters during January 1, 2008, to September 30, 2015, and October 1, 2015, to December 31, 2017, respectively.

To determine which method provides an optimal model, we built models identifying CIN2+ events using 3 distinct algorithms. The first algorithm was a prespecified set of CIN2+ tissue diagnosis codes alone, a method used by another claims-based study (17) classifying CIN2+ event status using ICD-9 codes for specific CIN2+ tissue diagnoses: 622.12 (CIN2) and 233.1 (CIN3). To identify events in the ICD-10 era, we mapped ICD-9 codes used by Flagg et al. (17) to corresponding ICD-10 codes (N87.1, N87.2, D06.0, D06.1, D06.7, and D06.9). The second algorithm was least absolute shrinkage and selection operator (LASSO) logistic regression, a machine learning method to build parsimonious models from correlated predictors by simultaneously conducting variable selection and regularization (18). The third algorithm was random forest classifiers, a machine learning method that averages several bootstrapped decision trees with various predictors and cutoff values to reduce overfitting (19). We conducted parameter tuning using a randomized search method, testing various combinations of number of trees, maximum predictor selection methods, maximum tree depths, minimum number of samples for a split, and minimum number of samples in a leaf node. The final parameters were selected based on highest mean validation score.

Models derived from LASSO and random forest algorithms were built using ICD-9 and ICD-10 training sets combined, creating a uniform model across eras. Correlation matrices were built in R (R core team, Vienna, Austria). LASSO was trained in Stata 16 (StataCorp, College Station, TX). Random forest was trained in Python 3.7.4 using the RandomForestClassifiers function from the scikit-learn package.

Statistical Analysis

We examined bivariate associations of demographic and coding characteristics of cervical diagnostic procedures between ICD eras using 2-sided Pearson's χ^2 tests. To compare concordance between each model-building methodology, we calculated percent agreement and Cohen kappa statistics. For bivariate and concordance tests, P values less than .05 were considered statistically significant.

Confusion matrices determined true positives, false positives, false negatives, and true negatives. To assess apparent validity, defined as model performance among samples used to develop the models (20), we examined discrimination and calibration of LASSO and random forest models among training sets by era. Discrimination was assessed using 6 performance measures: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and C-index. Calibration was assessed using calibration plots. Apparent validity was not assessed for CIN2+ tissue diagnosis codes alone because this method was not trained. We also examined discrimination and calibration in testing sets by era. Because CIN2+ diagnosis trends differ across ages, we also assessed model performance by age group (18-24 years, 25-29 years, 30-39 years) and ICD era among testing sets.

Binomial 95% confidence intervals (CIs) were calculated for all 6 performance measures using the Clopper-Pearson Exact method to test for statistically significant differences in model performance between the ICD eras. We assessed generalizability by comparing performance in training and testing sets by model and era; comparisons were considered statistically significant if 95% confidence intervals did not overlap. Lastly, we counted annual index events classified by each model to compare with HPV-IMPACT's confirmed annual number of incident CIN2+ events in the population.

Additional Analyses

The study (17) on which we based our first model (CIN2+ tissue diagnosis codes alone) was restricted to women screened for cervical cancer. To assess differences between women with cervical diagnostic procedures (our study population) vs those with screening tests (study population of Flagg et al. [17]), we replicated our methods among a cohort of women with cervical screening tests; however, when screening codes were used as the inclusion criteria, nearly one-half (46%) of confirmed CIN2+ events in the population were not captured (Supplementary Figure 1, available online). Therefore, this article reports results among women with cervical diagnostic procedures only.

Results

Cervical Diagnostic Procedure Characteristics

We identified 5639 TennCare-enrolled women aged 18-39 years residing in Davidson County, Tennessee, with 8549 (ICD-

9=7105; ICD-10=1444) cervical diagnostic procedures from 2008 to 2017 (Table 1). In the ICD-9 era, 885 of 7105 (12.5%) confirmed CIN2+ events occurred among women with cervical diagnostic procedures compared with 231 of 1444 (16.0%) in the ICD-10 era ($P < .001$). Compared with the ICD-9 era, a greater proportion of women with diagnostic procedures in the ICD-10 era were aged 30-39 years (ICD-9=27.6% vs ICD-10=46.8%) and other or unknown race or ethnicity (ICD-9=38.8% vs ICD-10=47.8%) ($P < .001$). The proportion of codes used in the ICD-9 vs ICD-10 era statistically significantly differed for the following code groupings: CIN2+ tissue diagnosis, nonspecific CIN tissue diagnosis, ASCUS diagnosis, HPV screening test, Pap smear or test, and cervical or vaginal biopsy ($P < .05$).

Model-Building Results

Models were trained using 60% of the total 8549 encounters, resulting in 5129 encounters (ICD-9=4263; ICD-10=866). Among the training set ($N=5129$), LASSO selected all code groupings as strong independent predictors of CIN2+; the strongest individual predictor was having a CIN2+ tissue diagnosis (Table 2, $\beta=5.34$). Other positive predictors included a nonspecific CIN diagnosis, HGSIL cytologic diagnosis, low-grade squamous intraepithelial lesion cytologic diagnosis, ASCUS diagnosis, cervical treatment procedure, or cervical or vaginal biopsy. Negative predictors included a CIN1 tissue diagnosis, HPV screening test, Pap smear or test, and HPV DNA test. Individual predictors were not highly correlated with one another; correlation coefficients were between -0.2 and 0.5 (Supplementary Figure 2, available online).

Optimal parameters for the random forest model included 23 trees, an automatic maximum predictor selection method, 36 maximum tree depth, 5 minimum samples for a split, and 8 minimum samples in a leaf node (Supplementary Table 2, available online). In the random forest model, having a CIN2+ tissue diagnosis code was the strongest predictor of CIN2+ event status (Table 2, importance score = 0.70).

Model Performance

The testing set included 40% of the total 8549 encounters, resulting in 3420 encounters (ICD-9=2842; ICD-10=578) (Supplementary Figure 3, available online). Classification agreement between models ranged from 92% to 98% (kappa range=0.74-0.91), with the highest concordance between LASSO and random forest.

Among the ICD-9 testing set ($N=2842$), 356 encounters were confirmed events (Figure 2). In the ICD-9 era, CIN2+ tissue diagnosis codes alone classified 624 encounters as CIN2+ events, of which 342 were correctly classified. Among the ICD-10 testing set ($N=578$), 90 encounters were confirmed events. In the ICD-10 era, CIN2+ tissue diagnosis codes alone classified 160 cervical diagnostic procedures as CIN2+ events, of which 88 were correctly classified. CIN2+ tissue diagnosis codes alone performed similarly between ICD-9 and ICD-10 eras: 96.1% (95% CI=93.5% to 97.8%) vs 97.8% (95% CI=92.2% to 99.7%) sensitivity, 88.7% (95% CI=87.3% to 89.9%) vs 85.3% (95% CI=81.8% to 88.3%) specificity, 54.8% (95% CI=50.8% to 58.8%) vs 55.0% (95% CI=46.9% to 62.9%) PPV, 99.4% (95% CI=98.9% to 99.7%) vs 99.5% (95% CI=98.3% to 99.9%) NPV, 89.6% (95% CI=88.4% to 90.7%) vs 87.2% (95% CI=84.2% to 89.8%) accuracy, and C-indices of 92.4% (95% CI=91.2% to 93.6%) vs 91.5% (95% CI=89.3% to

93.7%), respectively (Table 3, 95% confidence intervals overlapped).

Performance between training and testing sets for the model developed by LASSO was similar in both eras (Table 3, 95% confidence intervals overlapped). All LASSO performance measures in the testing set were similar between ICD-9 and ICD-10 eras: 77.3% (95% CI=72.5% to 81.5%) vs 81.1% (95% CI=71.5% to 88.6%) sensitivity, 93.0% (95% CI=91.9% to 94.0%) vs 90.2% (95% CI=87.2% to 92.7%) specificity, 61.3% (95% CI=56.6% to 65.8%) vs 60.3% (95% CI=51.0% to 69.1%) PPV, 96.6% (95% CI=95.8% to 97.3%) vs 96.3% (95% CI=94.1% to 97.8%) NPV, 91.0% (95% CI=89.9% to 92.1%) vs 88.8% (95% CI=85.9% to 91.2%), accuracy, and C-indices of 85.1% (95% CI=82.9% to 87.4%) vs 85.6% (95% CI=81.4% to 89.9%), respectively (95% confidence intervals overlapped). LASSO was well calibrated in both ICD eras and testing and training sets; expected and observed probabilities were similar (Supplementary Figure 4, available online).

Performance of the model developed by random forest in testing sets of ICD-9 and ICD-10 eras was similar: 70.2% (95% CI=65.2% to 74.9%) vs 75.6% (95% CI=65.4% to 84.0%) sensitivity, 93.8% (95% CI=92.8% to 94.8%) vs 90.6% (95% CI=87.6% to 93.0%) specificity, 62.0% (95% CI=57.1% to 66.8%) vs 59.6% (95% CI=50.1% to 68.7%) PPV, 95.7% (95% CI=94.8% to 96.4%) vs 95.3% (95% CI=92.9% to 97.0%) NPV, 90.9% (95% CI=89.8% to 91.9%) vs 88.2% (95% CI=85.3% to 90.7%) accuracy, and C-indices of 82.0% (95% CI=79.6% to 84.5%) vs 83.1% (95% CI=78.4% to 87.7%), respectively (Table 3, 95% confidence intervals overlapped). However, this model was not generalizable in the ICD-9 era, with statistically significant differences in sensitivity, NPV, accuracy, and C-index between training and testing sets (95% confidence intervals did not overlap). Random forest was well-calibrated for both eras and testing and training sets (Supplementary Figure 4, available online).

Model Performance by Age Group

Performance of CIN2+ tissue diagnosis codes alone was similar between ICD-9 and ICD-10 eras across all age groups, except for C-index among those aged 25-29 years (Table 4). C-indices statistically significantly differed between ages 18-24 years (94.6%, 95% CI=93.4% to 95.8%) and 25-29 years (98.7%, 95% CI=97.5% to 99.5%) as well as 25-29 years (98.7%, 95% CI=97.5% to 99.5%) and 30-39 years (91.9%, 95% CI=89.7% to 94.1%) in the ICD-9 era for CIN2+ tissue diagnosis codes alone (95% confidence intervals did not overlap). Models performed similarly between ICD-9 and ICD-10 eras across all age groups for LASSO and random forest. When comparing between age groups across ICD eras, all measures were similar for LASSO. For random forest, NPV statistically significantly differed between women aged 18-24 years (97.4%, 95% CI=96.3% to 98.3%) vs 25-29 years (94.2%, 95% CI=92.4% to 95.9%) in the ICD-9 era (95% confidence intervals did not overlap).

Estimation of Incident CIN2+ Events

From 2008 to 2017, HPV-IMPACT identified 983 incident CIN2+ events among TennCare-enrolled women aged 18-39 years residing in Davidson County (Figure 3; Supplementary Table 3, available online). When counting model-identified incident events compared with HPV-IMPACT's confirmed incident events, all claims-based models showed declining trends in CIN2+ incidence from 2008 to 2017, with some yearly classification variation. LASSO ($n=1007$) and random forest ($n=957$)

Table 1. Characteristics of cervical diagnostic procedures among TennCare-enrolled women aged 18-39 years in Davidson County, Tennessee, by ICD era^a

Characteristic	ICD-9 era N = 7105 No. (%)	ICD-10 era N = 1444 No. (%)	P
Confirmed CIN2+ ^b event			<.001
Yes	885 (12.5)	231 (16.0)	
No	6220 (87.5)	1213 (84.0)	
Age group, y			<.001
18-24	3062 (43.1)	261 (18.1)	
25-29	2081 (29.3)	507 (35.1)	
30-39	1962 (27.6)	676 (46.8)	
Race or ethnicity			<.001
NH White	2011 (28.3)	341 (23.6)	
NH Black	2184 (30.7)	382 (26.5)	
NH other or unknown	2755 (38.8)	690 (47.8)	
Hispanic	155 (2.2)	31 (2.2)	
CIN2+ ^b tissue diagnosis code			<.001
Yes	1508 (21.2)	381 (26.4)	
No	5597 (78.8)	1063 (73.6)	
Nonspecific CIN tissue diagnosis code			<.001
Yes	808 (11.4)	119 (8.2)	
No	6297 (88.6)	1325 (91.8)	
HGSIL cytologic diagnosis code			.20
Yes	845 (11.9)	189 (13.1)	
No	6260 (88.1)	1255 (86.9)	
CIN1 tissue diagnosis code			.95
Yes	1831 (25.8)	371 (25.7)	
No	5274 (74.2)	1073 (74.3)	
Low-grade squamous intraepithelial lesion cytologic diagnosis code			.44
Yes	2492 (35.1)	491 (34.0)	
No	4613 (64.9)	953 (66.0)	
ASCUS diagnosis code			.03
Yes	2480 (34.9)	547 (37.9)	
No	4625 (65.1)	897 (62.1)	
HPV screening test code			<.001
Yes	173 (2.4)	167 (11.6)	
No	6932 (97.6)	1277 (88.4)	
Pap smear or test code			.02
Yes	4987 (70.2)	1059 (73.3)	
No	2118 (29.8)	385 (26.7)	
HPV DNA test code			.11
Yes	3434 (48.3)	731 (50.6)	
No	3671 (51.7)	713 (49.4)	
Cervical treatment procedure code			.65
Yes	469 (6.6)	100 (6.9)	
No	6636 (93.4)	1344 (93.1)	
Cervical or vaginal biopsy code			<.001
Yes	3140 (44.2)	735 (50.9)	
No	3965 (55.8)	709 (49.1)	

^aThe ICD-9 era includes procedures from January 1, 2008, through September 30, 2015; the ICD-10 era includes procedures from October 1, 2015, through December 31, 2017. ASCUS = atypical squamous cells of undetermined significance; CIN = cervical intraepithelial neoplasia; DNA = deoxyribonucleic acid; HGSIL = high-grade squamous intraepithelial lesion; HPV = human papillomavirus; ICD = International Classification of Diseases, Clinical Modification; NH = non-Hispanic; Pap = Papanicolaou; TennCare = Tennessee Medicaid.

^bCIN2+ includes CIN2, CIN3, and adenocarcinoma in situ.

more closely captured the true number of population HPV-IMPACT incident events (n = 983) compared with CIN2+ tissue diagnosis codes alone (n = 1245).

Discussion

We validated claims-based models for estimating the number of CIN2+ events in ICD-9 and ICD-10 eras, which performed well and are optimized for public health surveillance and trend analyses. Among women with cervical diagnostic procedures, the

LASSO model most closely identified the population's confirmed incident CIN2+ events, with no statistically significant differences in performance between ICD eras. Because LASSO and random forest performed comparably well, model averaging could be an acceptable method; however, LASSO was more internally generalizable, with no statistically significant differences in performance between testing and training sets in both eras. Further, LASSO may be easier to understand and replicate within other databases compared with random forest because LASSO is a linear model.

Table 2. Beta coefficients and predictor importance scores of LASSO and random forest algorithms^a to classify CIN2+^b event status in the training set (N = 5129) of cervical diagnostic procedures among TennCare-enrolled women aged 18-39 years in Davidson County, Tennessee

Predictors	LASSO beta coefficients	Random forest predictor importance scores
Constant	-5.915605	—
CIN2+ tissue diagnosis	5.341873	0.695894
Cervical treatment procedure	0.9440706	0.089150
Cervical or vaginal biopsy	0.9414902	0.032999
High-grade squamous intraepithelial lesion diagnosis	0.9338596	0.095700
Nonspecific CIN diagnosis	0.3964537	0.028032
Low-grade squamous intraepithelial lesion diagnosis	0.3541705	0.010605
ASCUS diagnosis	0.2838765	0.010486
CIN1 tissue diagnosis	-0.2115674	0.015590
HPV DNA test	-0.2082338	0.008846
Pap smear or test	-0.1695168	0.011962
HPV screening test	-0.0893877	0.000737

^aLASSO and random forest algorithms were built using training sets of both ICD-9 and ICD-10 eras combined. ASCUS = atypical squamous cells of undetermined significance; CIN = cervical intraepithelial neoplasia; DNA = deoxyribonucleic acid; HPV = human papillomavirus; ICD = International Classification of Diseases, Clinical Modification; LASSO = least absolute shrinkage and selection operator; Pap = Papanicolaou; TennCare = Tennessee Medicaid.

^bCIN2+ includes CIN2, CIN3, and adenocarcinoma in situ.

Legend

	Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)
Classified Events (n)	True Positives	False Positives	Total Classified Events
Classified Non-Events (n)	False Negatives	True Negatives	Total Classified Non-Events
Total Gold Standard (N)	Total Confirmed Events	Total Confirmed Non-Events	Total Sample Size

CIN2+ Tissue Diagnosis Codes Alone	ICD-9 Era			ICD-10 Era				
		Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)		Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)
	Classified Events (n)	342	282	624	Classified Events (n)	88	72	160
	Classified Non-Events (n)	14	2,204	2,218	Classified Non-Events (n)	2	416	418
	Total Gold Standard (N)	356	2,486	2,842	Total Gold Standard (N)	90	488	578

LASSO	ICD-9 Era			ICD-10 Era				
		Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)		Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)
	Classified Events (n)	275	174	449	Classified Events (n)	73	48	121
	Classified Non-Events (n)	81	2,312	2,393	Classified Non-Events (n)	17	440	457
	Total Gold Standard (N)	356	2,486	2,842	Total Gold Standard (N)	90	488	578

Random Forest	ICD-9 Era			ICD-10 Era				
		Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)		Confirmed Events (n)	Confirmed Non-Events (n)	Total Classified (N)
	Classified Events (n)	250	153	403	Classified Events (n)	68	46	114
	Classified Non-Events (n)	106	2,333	2,439	Classified Non-Events (n)	22	442	464
	Total Gold Standard (N)	356	2,486	2,842	Total Gold Standard (N)	90	488	578

Figure 2. Confusion matrices of claims-based models to classify cervical intraepithelial neoplasia (CIN) grades 2+ (CIN2+) event status in the testing set of cervical diagnostic procedures among Tennessee Medicaid (TennCare)-enrolled women aged 18-39 years in Davidson County, Tennessee, by International Classification of Diseases, Clinical Modification (ICD) era. CIN2+ includes CIN2, CIN3, and adenocarcinoma in situ. The ICD-9 era includes procedures from January 1, 2008, through September 30, 2015; the ICD-10 era includes procedures from October 1, 2015, through December 31, 2017. LASSO = least absolute shrinkage and selection operator.

Table 3. Performance of claims-based models to classify CIN2+^a event status among cervical diagnostic procedures of TennCare-enrolled women aged 18-39 years in Davidson County, Tennessee, by ICD era^b

Performance Measure	CIN2+ tissue diagnosis codes alone						LASSO						Random forest								
	ICD-9 (N = 7105)		ICD-10 (N = 1444)		ICD-9 (N = 7105)		ICD-10 (N = 1444)		ICD-9 (N = 7105)		ICD-10 (N = 1444)		ICD-9 (N = 7105)		ICD-10 (N = 1444)		ICD-9 (N = 7105)		ICD-10 (N = 1444)		
	Testing set (n = 2842)	Training set (n = 4263)	Testing set (n = 2842)	Training set (n = 866)	Testing set (n = 2842)	Training set (n = 866)	Testing set (n = 2842)	Training set (n = 866)	Testing set (n = 2842)	Training set (n = 4263)	Testing set (n = 2842)	Testing set (n = 578)	Training set (n = 866)	Testing set (n = 2842)	Training set (n = 4263)	Testing set (n = 2842)	Testing set (n = 578)	Training set (n = 866)	Testing set (n = 2842)	Training set (n = 4263)	Testing set (n = 2842)
Sensitivity, % (95% CI)	96.1 (93.5 to 97.8)	97.8 (92.2 to 99.7)	97.3 (72.5 to 81.5)	82.0 (78.5 to 85.2)	77.3 (72.5 to 81.5)	75.2 (67.2 to 82.1)	81.1 (71.5 to 88.6)	79.8 ^c (76.1 to 83.1)	70.2 ^c (65.2 to 74.9)	79.8 ^c (76.1 to 83.1)	70.2 ^c (65.2 to 74.9)	81.1 (71.5 to 88.6)	75.2 (67.2 to 82.1)	77.3 (72.5 to 81.5)	79.8 ^c (76.1 to 83.1)	70.2 ^c (65.2 to 74.9)	81.1 (71.5 to 88.6)	75.9 (68.0 to 82.7)	70.2 ^c (65.2 to 74.9)	79.8 ^c (76.1 to 83.1)	70.2 ^c (65.2 to 74.9)
Specificity, % (95% CI)	88.7 (87.3 to 89.9)	85.3 (81.8 to 88.3)	93.0 (91.9 to 94.0)	94.2 (93.4 to 94.9)	93.0 (91.9 to 94.0)	93.1 (91.0 to 94.8)	90.2 (87.2 to 92.7)	95.0 (94.3 to 95.7)	93.8 (92.8 to 94.8)	95.0 (94.3 to 95.7)	93.8 (92.8 to 94.8)	90.2 (87.2 to 92.7)	93.1 (91.0 to 94.8)	93.0 (91.9 to 94.0)	95.0 (94.3 to 95.7)	93.8 (92.8 to 94.8)	90.2 (87.2 to 92.7)	93.5 (91.5 to 95.2)	93.8 (92.8 to 94.8)	95.0 (94.3 to 95.7)	93.5 (91.5 to 95.2)
PPV, % (95% CI)	54.8 (50.8 to 58.8)	55.0 (46.9 to 62.9)	61.3 (56.6 to 65.8)	66.8 (63.0 to 70.4)	61.3 (56.6 to 65.8)	68.0 (60.0 to 75.2)	60.3 (51.0 to 69.1)	69.5 (65.7 to 73.2)	62.0 (57.1 to 66.8)	69.5 (65.7 to 73.2)	62.0 (57.1 to 66.8)	60.3 (51.0 to 69.1)	68.0 (60.0 to 75.2)	61.3 (56.6 to 65.8)	69.5 (65.7 to 73.2)	62.0 (57.1 to 66.8)	60.3 (51.0 to 69.1)	69.5 (61.6 to 76.6)	62.0 (57.1 to 66.8)	69.5 (65.7 to 73.2)	62.0 (57.1 to 66.8)
NPV, % (95% CI)	99.4 (98.9 to 99.7)	99.5 (98.3 to 99.9)	96.6 (95.8 to 97.3)	97.4 (96.8 to 97.9)	96.6 (95.8 to 97.3)	95.1 (93.2 to 96.5)	96.3 (94.1 to 97.8)	97.1 ^c (96.5 to 97.6)	95.7 ^c (94.8 to 96.4)	97.1 ^c (96.5 to 97.6)	95.7 ^c (94.8 to 96.4)	96.3 (94.1 to 97.8)	95.1 (93.2 to 96.5)	96.6 (95.8 to 97.3)	97.1 ^c (96.5 to 97.6)	95.7 ^c (94.8 to 96.4)	96.3 (94.1 to 97.8)	95.2 (93.4 to 96.7)	95.7 ^c (94.8 to 96.4)	97.1 ^c (96.5 to 97.6)	95.2 (93.4 to 96.7)
Accuracy, % (95% CI)	89.6 (88.4 to 90.7)	87.2 (84.2 to 89.8)	91.0 (89.9 to 92.1)	92.7 (91.9 to 93.5)	91.0 (89.9 to 92.1)	90.2 (88.0 to 92.1)	88.8 (85.9 to 91.2)	93.2 ^c (92.4 to 93.9)	90.9 ^c (89.8 to 91.9)	93.2 ^c (92.4 to 93.9)	90.9 ^c (89.8 to 91.9)	88.8 (85.9 to 91.2)	90.2 (88.0 to 92.1)	91.0 (89.9 to 92.1)	93.2 ^c (92.4 to 93.9)	90.9 ^c (89.8 to 91.9)	88.8 (85.9 to 91.2)	90.6 (88.5 to 92.5)	90.9 ^c (89.8 to 91.9)	93.2 ^c (92.4 to 93.9)	90.6 (88.5 to 92.5)
C-Index, % (95% CI)	92.4 (91.2 to 93.6)	91.5 (89.3 to 93.7)	85.1 (82.9 to 87.4)	88.1 (86.5 to 89.8)	85.1 (82.9 to 87.4)	84.1 (80.5 to 87.8)	85.6 (81.4 to 89.9)	87.4 ^c (85.7 to 89.2)	82.0 ^c (79.6 to 84.5)	87.4 ^c (85.7 to 89.2)	82.0 ^c (79.6 to 84.5)	85.6 (81.4 to 89.9)	84.1 (80.5 to 87.8)	85.1 (82.9 to 87.4)	87.4 ^c (85.7 to 89.2)	82.0 ^c (79.6 to 84.5)	85.6 (81.4 to 89.9)	84.7 (81.1 to 88.4)	82.0 ^c (79.6 to 84.5)	87.4 ^c (85.7 to 89.2)	84.7 (81.1 to 88.4)

^aCIN2+ includes CIN2, CIN3, and adenocarcinoma in situ. CI = confidence interval; CIN = cervical intraepithelial neoplasia; ICD = International Classification of Diseases, clinical modification; LASSO = least absolute shrinkage and selection operator; NPV = negative predictive value; PPV = positive predictive value; TennCare = Tennessee Medicaid.

^bThe ICD-9 era includes procedures from January 1, 2008, through September 30, 2015; the ICD-10 era includes procedures from October 1, 2015, through December 31, 2017.

^cPerformance between the training and testing sets are statistically significantly different (95% confidence intervals do not overlap with each other).

Table 4. Performance of claims-based models to classify CIN2+^a event status by age group in the testing set of cervical diagnostic procedures among TennCare-enrolled women aged 18-39 years in Davidson County, Tennessee, by ICD era^b

Performance measure	Aged 18-24 y (N = 1349)		Aged 25-29 y (N = 1033)		Aged 30-39 y (N = 1038)	
	ICD-9 (n = 1248)	ICD-10 (n = 101)	ICD-9 (n = 835)	ICD-10 (n = 198)	ICD-9 (n = 759)	ICD-10 (n = 279)
CIN2+ tissue diagnosis codes alone						
Sensitivity, % (95% CI)	99.2 (95.5 to 99.8)	91.7 (61.5 to 99.8)	93.0 (86.6 to 96.9)	100.0 (91.4 to 100.0)	95.8 (90.5 to 98.6)	97.3 (85.8 to 99.9)
Specificity, % (95% CI)	90.0 (88.1 to 91.7)	89.9 (81.7 to 95.3)	87.2 (84.6 to 89.6)	80.9 (73.9 to 86.7)	88.0 (85.2 to 90.4)	86.4 (81.4 to 90.4)
PPV, % (95% CI)	51.7 (45.1 to 58.3)	55.0 (31.5 to 76.9)	53.3 (46.3 to 60.6)	57.7 (45.4 to 69.4)	59.9 (52.6 to 66.9)	52.2 (39.8 to 64.4)
NPV, % (95% CI)	99.9 (99.5 to 100.0)	98.8 (93.3 to 100.0)	98.7 (97.5 to 99.5)	100.0 (97.1 to 100.0)	99.1 (98.0 to 99.7)	99.5 (97.4 to 100.0)
Accuracy, % (95% CI)	90.9 (89.1 to 92.4)	90.1 (82.5 to 95.2)	88.0 (85.6 to 90.2)	84.9 (79.1 to 89.5)	89.2 (86.8 to 91.3)	87.8 (83.4 to 91.4)
C-Index, % (95% CI)	94.6 (93.4 to 95.8) ^c	90.8 (82.0 to 99.5)	98.7 (97.5 to 99.5) ^{c,d}	90.5 (87.4 to 93.5) ^d	91.9 (89.7 to 94.1) ^c	91.8 (88.4 to 95.3)
LASSO						
Sensitivity, % (95% CI)	81.1 (73.1 to 87.7)	75.0 (42.8 to 94.5)	74.6 (65.6 to 82.3)	80.5 (65.1 to 91.2)	75.8 (67.2 to 83.2)	83.8 (68.0 to 93.8)
Specificity, % (95% CI)	93.7 (92.1 to 95.0)	94.4 (87.4 to 98.2)	92.0 (89.7 to 93.8)	87.9 (81.7 to 92.6)	93.0 (90.7 to 94.8)	90.1 (85.6 to 93.5)
PPV, % (95% CI)	58.2 (50.4 to 65.7)	64.3 (35.1 to 87.2)	59.4 (50.9 to 67.6)	63.5 (49.0 to 76.4)	66.9 (58.3 to 74.7)	56.4 (42.3 to 69.7)
NPV, % (95% CI)	97.9 (96.8 to 98.6)	96.6 (90.3 to 99.3)	95.8 (94.0 to 97.2)	94.5 (89.5 to 97.6)	95.3 (93.4 to 96.9)	97.3 (94.3 to 99.0)
Accuracy, % (95% CI)	92.5 (90.1 to 93.9)	92.1 (85.0 to 96.5)	89.6 (87.3 to 91.6)	86.4 (80.8 to 90.8)	90.3 (87.9 to 92.3)	89.3 (85.0 to 92.6)
C-Index, % (95% CI)	87.4 (83.9 to 91.0)	84.7 (71.7 to 97.7)	83.3 (79.1 to 87.4)	84.2 (77.5 to 90.9)	84.4 (80.4 to 88.4)	86.9 (80.6 to 93.2)
Random forest						
Sensitivity, % (95% CI)	77.0 (68.6 to 84.2)	50.0 (21.1 to 78.9)	64.9 (55.4 to 73.6)	73.2 (57.1 to 85.8)	68.3 (59.2 to 76.5)	86.5 (71.2 to 95.5)
Specificity, % (95% CI)	94.7 (93.2 to 95.9)	94.4 (87.4 to 98.2)	92.4 (90.2 to 94.2)	89.8 (84.0 to 94.1)	94.1 (91.9 to 95.8)	89.7 (85.1 to 93.2)
PPV, % (95% CI)	61.0 (52.9 to 68.8)	54.5 (23.4 to 83.3)	57.4 (48.4 to 66.0)	65.2 (49.8 to 78.6)	68.3 (59.2 to 76.5)	56.1 (42.4 to 69.3)
NPV, % (95% CI)	97.4 (96.3 to 98.3) ^c	93.3 (86.1 to 97.5)	94.3 (92.4 to 95.9) ^c	92.8 (87.4 to 96.3)	94.1 (91.9 to 95.8)	97.7 (94.8 to 99.3)
Accuracy, % (95% CI)	93.0 (91.4 to 94.3)	89.1 (81.4 to 94.4)	88.6 (86.3 to 90.7)	86.4 (80.8 to 90.8)	90.0 (87.6 to 92.0)	89.3 (85.0 to 92.6)
C-Index, % (95% CI)	85.9 (82.1 to 89.7)	72.2 (57.2 to 87.2)	78.6 (74.1 to 83.2)	81.5 (74.2 to 88.8)	81.2 (76.9 to 85.5)	88.1 (82.2 to 94.0)

^aCIN2+ includes CIN2, CIN3, and adenocarcinoma in situ. CIN = cervical intraepithelial neoplasia; ICD = International Classification of Diseases, Clinical Modification; LASSO = least absolute shrinkage and selection operator; NPV = negative predictive value; PPV = positive predictive value; TennCare = Tennessee Medicaid.

^bThe ICD-9 era includes procedures from January 1, 2008, through September 30, 2015; the ICD-10 era includes procedures from October 1, 2015, through December 31, 2017.

^cPerformance between age groups, either aged 18-24 years vs 25-29 years, 18-24 years vs 30-39 years, or 25-29 years vs 35-39 years are statistically significantly different (95% confidence intervals do not overlap with each other).

^dPerformance between the ICD-9 and ICD-10 eras are statistically significantly different (95% confidence intervals do not overlap with each other).

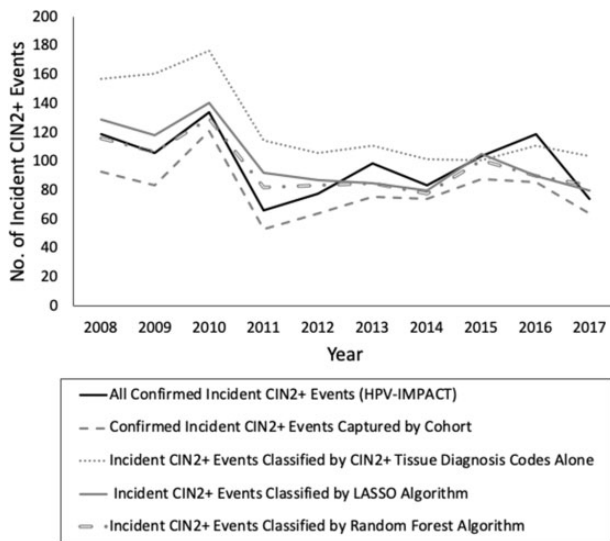


Figure 3. Annual number of incident cervical intraepithelial neoplasia (CIN) grades 2+ (CIN2+) events identified by claims-based models among Tennessee Medicaid (TennCare)-enrolled women aged 18-39 years residing in Davidson County, Tennessee, who had cervical diagnostic procedures from 2008 to 2017. CIN2+ includes CIN2, CIN3, and adenocarcinoma in situ; incident events were determined by applying each model to the cohort of cervical diagnostic procedures and counting index events classified by each model. HPV-IMPACT = human papillomavirus vaccine impact monitoring project; LASSO = least absolute shrinkage and selection operator.

When stratifying by age group, LASSO model performance was similar across age groups and eras. In our study population, the distribution of age groups differed across eras, which may be explained by epidemiologic shifts in disease occurrence and detection from changes in cervical screening guidelines and the impact of HPV vaccination over time. In 2012, updated guidelines recommended against cervical screening for women aged younger than 21 years (21), contributing to decreases in young women receiving Pap smears and cervical diagnostic procedures. Further, cervical biopsy data from the HPV-IMPACT monitoring project demonstrated declines in CIN2+ incidence among younger aged patients (18-24 years) who may have benefited from the HPV vaccine and increasing trends among older aged patients (30-39 years) from 2008 to 2015 (11). Due to differences in screening and CIN2+ trends across ages over time, changes in characteristics between eras seem reasonable.

Although using CIN2+ tissue diagnosis codes alone to identify CIN2+ events is intuitive, this approach had relatively low specificity, resulting in overestimated event classifications. We also observed overestimated incident CIN2+ event estimations after applying this approach to cervical diagnostic procedures and counting classified index events. When aiming for an accurate estimation of “true” population disease rates, specificity should be optimized to “rule-in” identified events. Although random forest had the highest specificity, this model was not as generalizable as LASSO. At the same time, from a data science perspective, it is important not to build a perfectly accurate model because this might mean the model is overtrained and is merely memorizing the training set’s data patterns, which would limit the external generalizability of the model.

One study (22) validated claims-based algorithms identifying high-grade cervical dysplasia, including HGSIL, CIN2, CIN3, and cervical cancer. However, the study was published before ICD-10; therefore, it included only ICD-9 codes, which is not useful

for assessing trends past 2015. Additionally, events were identified based on sets of rules (eg, at least 1-2 diagnosis codes and 1-2 procedure codes) then confirmed using chart analyses from a linked electronic health system. Thus, the algorithm measured PPV and not sensitivity; many true events may have been missed by the inclusion criteria. The study was restricted to women with abnormal Pap test codes, which we found to exclude nearly one-half of true events within our study population.

Claims data come with limitations. We were unable to test model performance by demographic sectors besides age group, such as race or ethnicity because 40% of our sample self-reported their race or ethnicity as other or unknown, with increases in this classification over time due to increasing proportions of enrollees not identifying in a single racial group (23,24). Because our sample was limited to women with qualifying diagnostic procedural codes, only 75% of HPV-IMPACT’s confirmed events were captured. Missingness may be because some women were retroactively enrolled in TennCare around the time of diagnosis and procedure codes were not captured, codes may have been nonspecific, or procedures were not billed. Whether these issues would apply to other insurance databases is unknown. We were unable to validate outside of Davidson County, Tennessee; thus, model performance in populations with different CIN2+ prevalence or demographics, such as in non-Medicaid populations, may differ and should be examined in future studies.

Our study had notable strengths. We used gold-standard data from HPV-IMPACT, which underwent extensive audits to ensure high-quality data. Because these data were population based, we could build models optimized for surveillance and trend analyses, prioritizing the estimation of population CIN2+ incidence. Additionally, we built models using machine learning methods, gaining valuable information from each method. To our knowledge, our study is the first to validate claims-based models for identifying incident CIN2+ events in ICD-9 and ICD-10 eras. Although we did not have access to external data to validate our models in outside populations, we were still able to demonstrate good internal generalizability between training and testing sets from our own sample. A potential next step could be an external validation of our models in another HPV-IMPACT partnering site to test how well these models perform in a different population.

Examining CIN2+ incident trends after 2015 is valuable for evaluating the HPV vaccine’s impact on reducing cervical precancers. These ecologic analyses are important because the vaccine has both direct effects (on vaccinated persons) and indirect effects (on those exposed to vaccinated persons) that are not captured in traditional vaccine effectiveness analyses. Since the vaccine’s introduction in 2006 (25), assessing US trends in CIN2+ incidence has been limited to populations with adequate cervical biopsy data (2, 8-13). Claims-based studies without access to population-based cervical biopsies are limited by the 2015 ICD-10 transition (17). Our study bridges these gaps by developing a simple model that may be uniformly applied to ICD-9 and ICD-10 eras with similar performance to assess more recent CIN2+ trends. This study expands options for CIN2+ surveillance by providing an alternate metric for identifying CIN2+ events in populations where cervical biopsy data are unavailable.

Funding

This work was supported by the National Institutes of Health (5TL1TR002244, R01CA207401) and the Emerging

Infections Cooperative Agreement from the Centers for Disease Control and Prevention (5U01C10003).

Notes

Role of the funder: The funders had no role in the design of the study; collection, management, analysis, and interpretation of the data; writing, review, or approval of the manuscript; and the decision to submit for publication.

Disclosures: The authors have no conflicts to disclose. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health or the Centers for Disease Control and Prevention.

Author contributions: JZS, MRG, JCS, PCH: Study conceptualization, methodology, and design; JZS, LDN: Formal analysis and software coding; JZS: Original draft of manuscript; JZS, ABR: Visualization; JZS, MRG, LDN, JCS, EFM, MP, ABR, PCH: Critical review of manuscript and revisions; MRG, EFM, MP: Project administration and coordination of resources; MRG, PCH: Study supervision.

Acknowledgements: The authors acknowledge the Division of TennCare of the Tennessee Department of Finance and Administration for providing the data to the HPV-IMPACT Monitoring Project; Dr Ronald Alvarez, Vanderbilt University Medical Center, for providing medical expertise regarding CIN2+ and administrative codes; and Dr Charreau Bell, Vanderbilt University Data Science Institute, for providing methodologic and modeling advice.

Data Availability

The data underlying this article were provided to Vanderbilt University Medical Center (VUMC) by the Division of TennCare of the Tennessee Department of Finance and Administration under a contract that does not permit VUMC to share the data with external parties. Researchers may request data from the Division of TennCare of the Tennessee Department of Finance and Administration.

References

- Ylitalo N, Josefsson A, Melbye M, et al. A prospective study showing long-term infection with human papillomavirus 16 before the development of cervical carcinoma in situ. *Cancer Res*. 2000;60(21):6027–6032.
- McClung NM, Gargano JW, Park IU, et al. HPV-IMPACT Working Group. Estimated number of cases of high-grade cervical lesions diagnosed among women—United States, 2008 and 2016. *MMWR Morb Mortal Wkly Rep*. 2019;68(15):337–343.
- Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *Lancet*. 2007;370(9590):890–907.
- Henk HJ, Insinga RP, Singhal PK, Darkow T. Incidence and costs of cervical intraepithelial neoplasia in a US commercially insured population. *J Low Genit Tract Dis*. 2010;14(1):29–36.
- Novaes HMD, Itria A, Silva G. e, et al. Annual national direct and indirect cost estimates of the prevention and treatment of cervical cancer in Brazil. *Clinics (Sao Paulo)*. 2015;70(4):289–295.
- Östensson E, Silfverschiöld M, Greiff L, et al. The economic burden of human papillomavirus-related precancers and cancers in Sweden. *Plos One*. 2017;12(6):e0179520.
- Elam-Evans LD, Yankey D, Singleton JA, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13–17 years—United States, 2019. *MMWR Morb Mortal Wkly Rep*. 2020;69(33):1109–1116. doi: 10.15585/mmwr.mm6933a1.
- Oakley F, Desouki MM, Pemmaraju M, et al. Trends in high-grade cervical cancer precursors in the human papillomavirus vaccine era. *Am J Prev Med*. 2018;55(1):19–25.
- Benard VB, Castle PE, Jenison SA, et al.; for the New Mexico HPV Pap Registry Steering Committee. Population-based incidence rates of cervical intraepithelial neoplasia in the human papillomavirus vaccine era. *JAMA Oncol*. 2017;3(6):833–837.
- Niccolai LM, Julian PJ, Meek JI, McBride V, Hadler JL, Sosa LE. Declining rates of high-grade cervical lesions in young women in Connecticut, 2008–2011. *Cancer Epidemiol Biomarkers Prev*. 2013;22(8):1446–1450.
- Gargano JW, Park IU, Griffin MR, et al. Trends in high-grade cervical lesions and cervical cancer screening in five states, 2008–2015. *Clin Infect Dis*. 2019;68(8):1282–1291.
- Hairi S, Johnson ML, Bennett NM, et al.; HPV-IMPACT Working Group. Population-based trends in high-grade cervical lesions in the early human papillomavirus vaccine era in the United States. *Cancer*. 2015;121(16):2775–2781.
- McClung NM, Gargano JW, Bennett NM, et al. Trends in human papillomavirus vaccine types 16 and 18 in cervical precancers, 2008–2014. *Cancer Epidemiol Biomarkers Prev*. 2019;28(3):602–609.
- Cartwright DJ. ICD-9-CM to ICD-10-CM Codes. What? Why? How? *Adv Wound Care (New Rochelle)*. 2013;2(10):588–592.
- Centers for Disease Control and Prevention. International Classification of Diseases, (ICD-10-CM/PCS) Transition - Background. 2019. https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm. Accessed March 18, 2019.
- Centers for Disease Control and Prevention. Human papillomavirus vaccine impact monitoring project (HPV-IMPACT). 2018. <https://www.cdc.gov/ncird/surveillance/hpvimpact/overview.html>. Accessed February 9, 2019.
- Flagg EW, Torrone EA, Weinstock H. Ecological association of human papillomavirus vaccination with cervical dysplasia prevalence in the United States, 2007–2014. *Am J Public Health*. 2016;106(12):2211–2218.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodological)*. 1996;58(1):267–288.
- Tin Kam H. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Montreal, Quebec, Canada: Institute of Electrical and Electronics Engineers (IEEE), Vol. 1; 1995:278–282. doi: 10.1109/ICDAR.1995.598994.
- Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*. 2003;56(5):441–447.
- Saslow D, Solomon D, Lawson HW, et al.; ACS-ASCCP-ASCP Cervical Cancer Guideline Committee. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology Screening Guidelines for the Prevention and Early Detection of Cervical Cancer. *CA Cancer J Clin*. 2012;62(3):147–172.
- Kim SC, Gillet VG, Feldman S, et al. Validation of claims-based algorithms for identification of high-grade cervical dysplasia and cervical cancer. *Pharmacoepidemiol Drug Saf*. 2013;22(11):1239–1244.
- Shing JZ, Hull PC, Zhu Y, et al. Trends in anogenital wart incidence among Tennessee Medicaid enrollees, 2006–2014: the impact of human papillomavirus vaccination. *Papillomavirus Res*. 2019;7:141–149. doi: 10.1016/j.pvr.2019.04.007.
- Guerino P, James C. Race, ethnicity, and language preference in the health insurance marketplaces 2017 open enrollment period. 2017. <https://www.cms.gov/About-CMS/Agency-Information/OMH/research-and-data/information-products/data-highlights/Race-Ethnicity-and-Language-Preference-in-the-Health-Insurance-Marketplace>. Accessed July 23, 2018.
- Markowitz LE, Dunne E, Sariaiya M, Lawson H, Chesson H, Unger E. Quadrivalent human papillomavirus vaccine: recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Morb Mortal Wkly Rep*. 2007;56(2):1–24.