Machine Learning and Real-World Data to Predict Lung Cancer Risk in Routine Care

Urmila Chandran[1,2], Jenna Reps[1], Robert Yang[2], Anil Vachani[3], Fabien Maldonado[4], and Iftekhar Kalsekar[2]

ABSTRACT

Background: This study used machine learning to develop a 3-year lung cancer risk prediction model with large real-world data in a mostly younger population.

Methods: Over 4.7 million individuals, aged 45 to 65 years with no history of any cancer or lung cancer screening, diagnostic, or treatment procedures, with an outpatient visit in 2013 were identified in Optum's de-identified Electronic Health Record (EHR) dataset. A least absolute shrinkage and selection operator model was fit using all available data in the 365 days prior. Temporal validation was assessed with recent data. External validation was assessed with data from Mercy Health Systems EHR and Optum's de-identified Clinformatics Data Mart Database. Racial inequities in model discrimination were assessed with xAUCs.

Results: The model AUC was 0.76. Top predictors included age, smoking, race, ethnicity, and diagnosis of chronic obstructive pulmonary disease. The model identified a high-risk group with lung cancer incidence 9 times the average cohort incidence, representing 10% of patients with lung cancer. Model performed well temporally and externally, while performance was reduced for Asians and Hispanics.

Conclusions: A high-dimensional model trained using big data identified a subset of patients with high lung cancer risk. The model demonstrated transportability to EHR and claims data, while underscoring the need to assess racial disparities when using machine learning methods.

Impact: This internally and externally validated real-world data-based lung cancer prediction model is available on an open-source platform for broad sharing and application. Model integration into an EHR system could minimize physician burden by automating identification of high-risk patients.

Introduction

Lung cancer is the leading cause of cancer mortality in the US, surpassing breast, prostate, and pancreatic cancers combined (1). Although there have been improvements in clinical outcomes as a result of diagnostic and therapeutic advances, the overall 5-year survival is only 22% with the majority of lung cancers still being diagnosed at advanced stage (2). One contributing factor is the low uptake of low-dose CT for lung cancer screening (~7%) nationally (3), an intervention associated with a 20% relative reduction in lung cancer-specific mortality in high-risk patients. Further, the majority of lung cancers are diagnosed in individuals that do not meet screening eligibility criteria (4). Despite the expanded screening eligibility with the revised United States Preventive Services Task Force (USPSTF) guidelines (5), individuals that do not meet screening criteria have a high risk of lung cancer (6), supporting the need for prediction tools that can facilitate the identification of high-risk individuals and the adoption of screening in clinical practice.

Most available lung cancer risk prediction models are based on clinically rich data from trials requiring administration of questionnaires to patients as many of the predictor variables are not generally captured in routine clinical practice. Access to comprehensive real-world data from electronic health records (EHR) presents a resource-efficient and complementary approach for prediction research. Although two published prediction models have leveraged large real-world data from EHR in community settings (7, 8), neither model evaluation included external validation.

According to the NCI Surveillance Epidemiology and End Results database, while the median age for lung cancer is 71 years, almost 30% of lung cancers are diagnosed in individuals younger than 65 years of age (9). Therefore, as a proof-of-concept, this study assessed feasibility in using machine learning to predict lung cancer risk over a 3-year period, and to identify a 'high-risk subset' among a relatively young cohort of patients 45 to 65 years of age, in a large EHR dataset. Although the incidence of lung cancer is lower in younger individuals, the model aimed to identify adults that comprise a high-risk group that could benefit from more detailed risk assessment. Focusing on a younger cohort was proposed as a preliminary assessment of how well a model based on EHR data could identify individuals with high lung cancer risk even in a population with relatively low incidence.

Materials and Methods

We followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines for prediction model development and validation (10).

Study design and population

A machine learning–based model to predict 3-year lung cancer risk was developed using EHR data from Optum's de-identified Electronic Health Record dataset (Optum EHR). This retrospective cohort study included adults 45 to 65 years of age with an outpatient visit during January 1, 2013 to December 31, 2013, and at least 365 days of continuous observation in the database prior to cohort entry. The index date was the first outpatient visit date in 2013 when all eligibility criteria were met. To ensure that the model was predicting new

[1]Johnson & Johnson Global Epidemiology, Titusville, New Jersey. [2]Lung Cancer Initiative, Johnson & Johnson, New Brunswick, New Jersey. [3]University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania. [4]Vanderbilt University, Nashville, Tennessee.

Corresponding Author: Urmila Chandran, Johnson & Johnson Global Epidemiology, 1125 Trenton Harbourton Rd, Titusville, NJ 08560. Phone: 609-737-2699; E-mail: uchandr1@its.jnj.com

Cancer Epidemiol Biomarkers Prev 2023;32:337–43

doi: 10.1158/1055-9965.EPI-22-0873

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

©2022 The Authors; Published by the American Association for Cancer Research

diagnoses of lung cancer, individuals with any prior cancer recorded on or before the index date (except for nonmelanoma skin cancer), any prior lung cancer-related screening (chest CT scan), diagnostic or confirmatory imaging (CT scan, PET scan), procedures (biopsy, bronchoscopy), and cancer treatment (surgery, radiation, ablation, systemic therapy), as well as any prior lung cancer-related symptoms (hemoptysis) were excluded from the cohort (see flow chart of cohort attrition and study design in Supplementary Fig. S1A and S1B). The decision to exclude any prior cancer was made to minimize the impact of surveillance bias in cancer survivors. Patients who had no longitudinal data beyond the index date were excluded. No restrictions were placed on smoking status.

Data sources

Optum EHR is a multidimensional database that integrates data on outpatient visits, diagnostic procedures, medications, lab results, hospitalizations, and patient outcomes primarily from different EHR platforms and health systems in the US, including both publicly and privately insured as well as uninsured patients. This EHR database was chosen for model training and development due to the capture of smoking status data, which is generally available as a separate variable in an EHR, rather than solely through diagnosis codes.

For external model validation, two different datasets were used to identify cohorts with the same eligibility criteria and study period: (i) The EHR database available from the Mercy Health Systems (Mercy EHR), which is an integrated health system in Midwestern US, headquartered in St. Louis, Missouri and (ii) Optum's de-identified Clinformatics Data Mart Database (Optum Claims), which is an adjudicated administrative claims database for members with private health insurance, who are fully insured in commercial plans or in administrative services. The Optum Claims dataset was included for external validation to assess the applicability of the model in non-EHR settings.

All three databases were standardized into the Observational Medical Outcomes Partnership Common Data Model, which enables standardization of data from disparate observational databases into a common format to facilitate standardized analytics and efficiencies in research (see https://ohdsi.github.io/CommonDataModel/). As the predictive model was developed with the aim to be reproducible, a standardized clinical vocabulary for medical conditions, Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), was used. Due to the existence of mappings between SNOMED-CT and other vocabularies, medical condition code sets using SNOMED-CT can be readily translated into other vocabularies such as ICD-9-CM or ICD-10-CM.

The use of Optum databases was reviewed by the New England Institutional Review Board (IRB) and was determined to be exempt from study-specific IRB approval. The Mercy EHR data were anonymized and de-identified and therefore the study did not require Mercy IRB oversight.

Outcome of interest

The outcome to be predicted was risk of new lung cancer starting 1 day until 3 years (1,095 days) after index. Due to the absence of pathologic confirmation of disease in the three databases, new lung cancer was defined as an individual having ≥ 2 ICD diagnosis codes for primary lung cancer (ICD-9 code 162.XX or ICD-10 code C34.XX) within 60 days. In the published literature of real-world data, various algorithms have been used to define lung cancer, ranging from one diagnosis code (8, 11) or combination of diagnosis codes and treatment (11–14). In the absence of a standardized algorithm to identify lung cancer in real-world data, requiring ≥2 diagnosis codes for primary lung cancer was considered a robust definition. This approach limits the issues of a highly sensitive definition of requiring only one diagnosis code that could be coded as a 'rule out diagnosis' and a highly specific definition that requires treatment and excludes patients without treatment information.

Candidate covariates

Candidate covariates were identified using data recorded in the EHR within 1 year prior to the index date (including on index). A 1-year interval was chosen for covariate selection as this provides an acceptable trade-off between model performance and interpretability (15). The following group of covariates were included:

Binary demographic covariates indicating the presence (value 1) or absence (value 0) for gender (male – yes/no), age groups (aged 45–49, aged 50–54, aged 55–59, aged 60–64, and aged 65), race (White, Black, Asian, unknown), and ethnicity (Hispanic, non-Hispanic, unknown)

Binary concept covariates indicating the presence (value 1) or absence (value 0) per diagnoses, drug ingredient, whether a measurement was recorded (labs or other), medical device, procedure, or observation recorded in the 1-year prior to index (including index). However, only diagnoses, drug ingredients, measurements, medical devices, procedures, and observations recorded for a minimum of 20 patients 1 year prior to index in the Optum EHR study population were included.

The number of visits a patient had recorded per visit type (e.g., outpatient or inpatient) in the prior 1 year.

Smoking status recorded as current, previous, never smoker or missing for the patient in the prior 1 year. If a patient had been recorded as both a current and previous smoker in the prior year, they were considered a current smoker. If a patient was recorded as both a previous and never smoker in the prior year, they were considered a previous smoker. Smoking status data are available in EHR, in addition to diagnosis codes for tobacco dependence disorder, while Optum Claims only had the latter.

The measurement covariates in the model represent the record of a measurement being taken rather than the value (as values are often missing). For example, the measurement 'pack years' means a pack years measurement was recorded (with or without a value). The model did not include measurement values because: (i) the values are often poorly captured and would cause a missing data issue and (ii) the measurement units are often not standardized.

Missing data

The covariate construction was done in a way such that each binary concept covariate corresponds to whether a patient had a code (e.g., SNOMED or rxnorm) recorded in the prior 365 days (to index), which was 1 if they had a record and 0 otherwise; hence no missing values. All continuous covariates were normalized (value-min)/max. Age and gender were always recorded and never missing. Race, ethnicity, and smoking categories could be unreported, but this was addressed by adding an "unknown" category.

Statistical analysis

Patients were classified as having the outcome if they met the definition for new lung cancer during the 3-year time-at-risk period, which started 1 day after index and ended the date the outcome definition was met, death date, last follow-up, or end of time-at-risk, whichever occurred first. The Optum EHR cohort was randomly split (stratified by outcome) into training and test datasets, whereby 75% of the data were used to train the model and remaining 25% was used to internally validate the model.

The analysis used a prediction model framework that has previously demonstrated reproducibility to facilitate model sharing and allow external validation (16, 17). Three prediction modeling approaches were assessed – regularized least absolute shrinkage and selection operator (LASSO) logistic regression, XGBoost (AUC = 0.69), and a deep learning approach (AUC = 0.71). A high-dimensional LASSO regression prediction model was pursued for main results, as it not only demonstrated the best performance in preliminary analyses, but also renders a parsimonious model when including a large number of features. The LASSO regression model has a single hyper-parameter that controls the level of regularization. A 3-fold cross validation was used on the training data to learn the optimal hyper-parameter value, which has been shown to be sufficient in previous research (18).

Model discrimination was assessed using ROC plots, and calibration was assessed using calibration plots. The area under the ROC (AUC) corresponds to the probability that a randomly selected person with the outcome is assigned a higher risk of the outcome by the model than a randomly selected person without the outcome. Because the AUC hides the low precision issue when the outcome is rare (19), the area under the precision-recall curve and predictive lift [positive predictive value (PPV) divided by observed risk of outcome in the study population] was calculated for the test set cohort and in external validation. These allow assessment of the model's utility and ability to identify subsets of patients that have higher risk than the overall cohort, based on a prediction threshold. Calibration plots provided a visual assessment to evaluate if the model's predicted risk matched the observed risk in all three datasets.

To evaluate the ability of the model to risk stratify the population, a survival graph of lung cancer cases across the study period was generated to visualize the occurrence of outcomes over time for different risk groups.

External and temporal validation analyses

The trained model was validated in Mercy EHR and Optum Claims to evaluate external model performance. The model was temporally validated in Optum EHR by using the same eligibility criteria but with an outpatient visit in 2014, 2015, 2016, and 2017 as the index to predict 3-year risk.

Race and ethnicity bias assessment

To determine if the model was systematically biased in estimating risk for any race and ethnicity subgroup, the AUCs within each race and ethnicity and xAUCs (pronounced cross-AUCs) across each race and ethnicity were examined (20). The xAUC uses the same calculation as the AUC but only includes a subset of the whole population. The xAUC is the probability that a randomly chosen patient from one subgroup (e.g., White race) with the outcome has a higher risk assigned by the model than a randomly chosen patient without the outcome, not in that subgroup. Low xAUC values suggest the model is assigning a higher risk less often. If a model is equally assigning risk across the two subgroups, then the xAUCs should be similar to the whole population AUC.

Data availability

Optum EHR and Optum Claims were licensed from Optum. Mercy EHR data were purchased from Mercy Health Systems. Hence, the data are not publicly available, but can be licensed or purchased from the database owners.

All analyses were conducted using open-source tools from Observational Health Data Sciences and Informatics (OHDSI) (21). OHDSI is a collaboration of researchers and data holders with an interest in developing best practices for studies using observational healthcare data. The full analysis source code to replicate this study, including the model developed for this analysis is available at https://github.com/ohdsi-studies/lungCancerPrognostic.

Results

A total of 4,777,606 patients were identified for cohort inclusion in the Optum EHR dataset for the model. Cohorts for external validation included 237,491 patients in the Mercy EHR dataset and 1,504,069 patients in the Optum Claims dataset. Age, gender, and race distributions across the three datasets are shown in **Table 1**. In all three datasets, approximately 0.1% of patients had lung cancer during the 3-year follow-up.

Table 1. Distribution of cohort size, patient demographics, and model internal and external validation.

	Optum EHR (2013 index)	Mercy EHR (2013 index) – Validation #1	Optum Claims (2013 index) – Validation #2
Cohort size	**4,777,606**	**237,491**	**1,504,069**
Age in years			
45–49	23%	21%	27%
50–54	27%	26%	27%
55–59	26%	25%	24%
60–65	24%	28%	22%
Female gender	60%	61%	52%
White race	82%	90%	75%
Hispanic ethnicity	4%	2%	9%
Smoking status[a]			
Current smokers	9.5%	16.0%	Not available
Former smokers	14.2%	17.7%	
Never smokers	16.6%	51.5%	
Unreported	59.7%	14.8%	
COPD diagnosis in 365 days prior to index	1.5%	0.75%	1.8%
Lung cancers, *n*	3661	260	1121
% of lung cancers during the 3-year time-at-risk	0.08%	0.11%	0.07%
AUC (95% CI)	0.76 (0.75–0.78)	0.81 (0.79–0.84)	0.72 (0.71–0.74)

[a]Smoking data in EHR are captured using smoking status variable (current smoker, former smoker, never smoker, unreported) and diagnosis codes for tobacco dependence disorder. In Optum Claims, smoking is captured through diagnosis codes with no variable for smoking status.

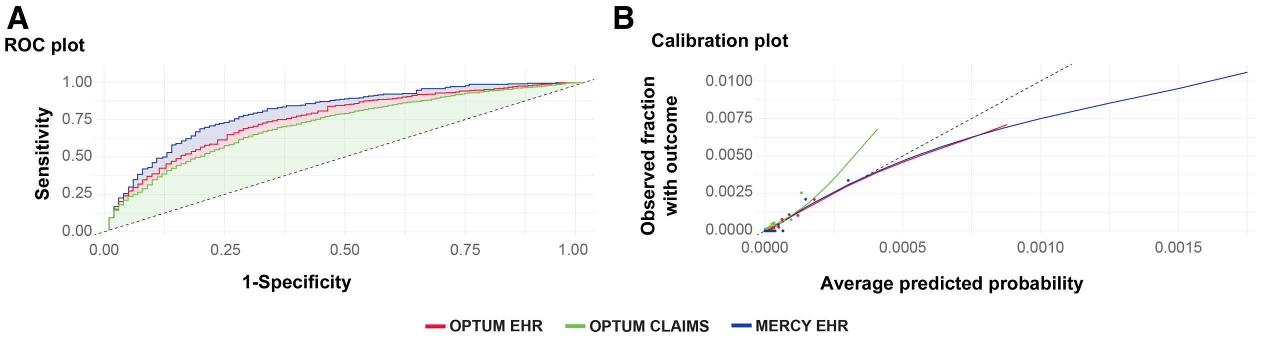

Figure 1.
ROC and calibration plots for internal and external validation. **A,** shows the area under the receiver operating characteristics curve for the main model trained in Optum EHR and externally validated on Mercy EHR and Optum Claims datasets. **B,** shows model calibration (in Optum EHR and in Mercy EHR and Optum Claims) as indicated by the observed fraction of patients in each decile and the corresponding mean predicted risk for each decile.

The model (Optum EHR) AUC was 0.76 with external validation AUCs of 0.81 in Mercy EHR and AUC 0.72 in Optum Claims. The high-dimensional model was fit with 16,633 candidate variables, with 278 variables with nonzero beta coefficients included in the final model. Covariates with some of the highest absolute coefficient values included being a current smoker, increasing age, ethnicity (not Hispanic), race (White), and a diagnosis of chronic obstructive pulmonary disease (COPD) in the prior year. Model performance was comparable within narrower age strata (Supplementary Table S1). The full model (Supplementary Table S2) can be accessed at: https://github.com/ohdsi-studies/lungCancerPrognostic/tree/master/inst/models/full_model.

Calibration and ROC plots for internal and external validation are shown in **Fig. 1**. The calibration plot demonstrates that the observed fraction of patients in each decile and the corresponding mean predicted risk for each decile (i.e., the ten dots) fall along the $x = y$ line, indicating good model calibration.

The model transported well to future data with AUCs of ≥ 0.76 in temporal validation analyses (**Table 2**). ROC and calibration plots for temporal validation are included in Supplementary Fig. S2.

Table 3 demonstrates the utility of the model. For example, if the model is used to identify patients at a prediction threshold of 0.48% (i.e., those with a risk ≥0.48% are considered high-risk), the PPV would be 0.72%, and we would identify 12,706 individuals (1.06% of the test set cohort) with 9.35 times the lung cancer incidence of the overall cohort and 10% of lung cancers (sensitivity) in the cohort. The model was able to maintain this ability of identifying high-risk patients that had approximately 9 times the average cohort risk in both Mercy EHR and Optum Claims datasets (**Table 3**), as well as with future years of data (Supplementary Table S3). The full range of operating characteristics for the model are shown in Supplementary Fig. S3.

The lung cancer occurrence survival plot in the target cohort over the 3-year time-at-risk in Optum EHR for different risk groups is shown in **Fig. 2**. The model showed ability to discriminate across

Table 2. Results from temporal validation in context with fitted model.

Cohort from Optum EHR data	Cohort sample	Lung cancer, *n* (%)	AUC (95% CI)
Cohort entry year 2013 (fitted model)	4,777,606	3,661 (0.08%)	0.76 (0.75–0.78)
Cohort entry year 2014	5,217,732	4360 (0.08%)	0.77 (0.76–0.78)
Cohort entry year 2015	5,568,584	4768 (0.09%)	0.78 (0.77–0.78)
Cohort entry year 2016	5,500,418	4569 (0.08%)	0.77 (0.77–0.78)
Cohort entry year 2017	5,434,897	4263 (0.08%)	0.77 (0.77–0.78)

Table 3. Metrics related to model utility.

	%PPV[a] at specified Sen				Predictive lift	No. screened	% screened
	Sen of 5%	Sen of 10%	Sen of 25%	Sen of 50%	at Sen of 10%	at Sen of 10%	at Sen of 10%
Optum EHR	0.94	0.72	0.43	0.24	9.35	12706[b]	1.06
Mercy EHR	1.34	0.97	0.73	0.45	8.85	2682	1.13
Optum Claims	0.94	0.68	0.32	0.19	9.06	16592	1.1

Abbreviation: Sen, sensitivity.
[a] PPV is defined as patients 'correctly' identified as lung cancer patient according to definition used [true positives] divided by patients identified by model to have lung cancer [true positives + false positives].
[b] 1.06% of patients in the test set cohort comprising 25% of overall sample.

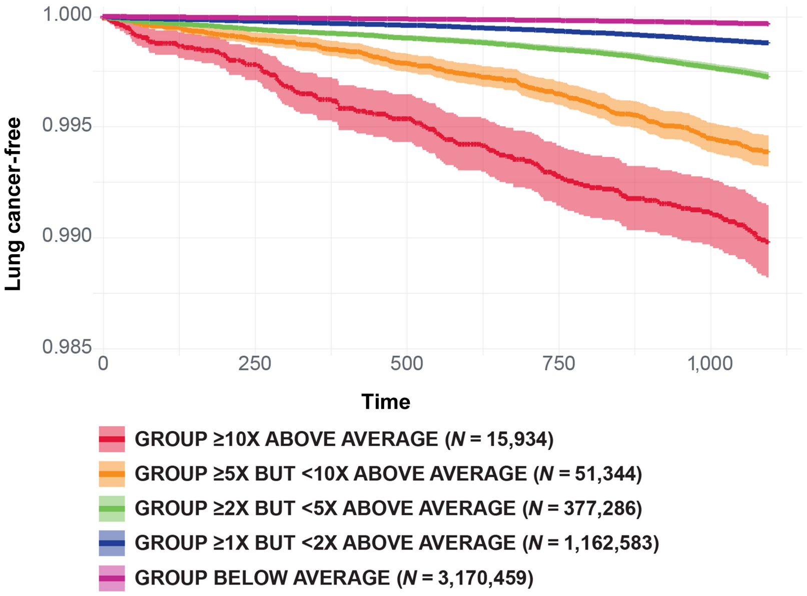

Figure 2.
Survival plot of lung cancer occurrence over 3 years (time in days) for different risk groups. This graph shows the occurrence of lung cancer cases in patient groups that are above or below average cohort risk during the 3-year time-at-risk period. The proportion of patients being undiagnosed with lung cancer is shown on the *Y*-axis and time in days on the *X*-axis.

various high- and low-risk groups that had over or below the average cohort risk, as demonstrated by the clear separation of lines, with the steepest drop in probability of being lung cancer–free for those with more than 5 times the average cohort incidence. The lack of case clustering close to the start of follow-up demonstrates that the lung cancers were likely new diagnoses and not prevalent. When the model was regenerated after excluding lung cancer cases in the first 180 days after index, model performance was similar [AUC, 0.76; 95% confidence interval (CI), 0.75–0.78].

The race- and ethnicity-stratified model AUCs as well as xAUCs in the race and ethnicity subgroups are shown in **Table 4**. Within subgroup AUCs appeared generally comparable for each race and ethnicity. However, xAUC analysis showed that while 85% of the time a randomly selected White patient with lung cancer in the 3-year time-at-risk has a higher risk assigned to them than a randomly selected non-White patient without lung cancer, a randomly selected Asian patient with lung cancer in the 3-year time-at-risk has a higher risk assigned to them than a randomly selected non-Asian patient without lung cancer only 54% of the time. Similarly, the model assigned a higher risk to a non-Hispanic lung cancer patient than a patient not in this subgroup without lung cancer 93% of the time, while a Hispanic patient with lung cancer was assigned a higher risk only 43% of the time than a patient not in the Hispanic subgroup without lung cancer. A simple model that only included age and smoking, similar to variables in the USPSTF criteria had an AUC of 0.72 with some improvement in xAUCs for Asians (xAUC: 0.60) and Hispanics (xAUC: 0.62).

Discussion

An EHR-based high-dimensional lung cancer risk prediction model developed using the largest sample to date was able to discriminate individuals with varying risk of lung cancer over a 3-year period. This model transported well to two external datasets and to future data. Along with known factors such as older age, smoking, and COPD, the model identified other covariates to provide a comprehensive prediction framework using structured EHR data.

Several lung cancer risk prediction models have been published in the literature, with some (e.g., PLCOm2012, LCDRAT) suggesting improved efficiency in identifying individuals for lung cancer screening compared with USPSTF criteria (22). All of these models were developed using comprehensive clinical data acquired in the context of longitudinal cohort studies or clinical trials. This results in models that are often difficult to translate to clinical practice given the lack of certain covariates in EHR data. In contrast, EHR-based models leverage large amounts of data collected as part of routine practice and can be developed in broader populations thereby serving as complementary tools to traditional prediction models. EHR models can be implemented with greater ease using automated approaches, which may result in greater efficiency for clinicians, while simultaneously reducing patient burden.

To our knowledge, there are only two other published EHR-based machine learning models (7, 8) of lung cancer risk prediction in the US, and neither included model validation on external datasets. The model from the Maine Health Information Exchange (8) predicted 1-year

Table 4. Results from race and ethnicity stratified models and xAUCs.

Race/Ethnicity	No. of lung cancers	AUC (95% CI) within race/ethnicity	xAUC (95% CI)
White	3206	0.76 (0.75–0.77)	0.85 (0.85–0.86)
Black or African American	297	0.77 (0.74–0.79)	0.75 (0.72–0.78)
Asian	29	0.70 (0.60–0.79)	0.54 (0.44–0.65)
Non-Hispanic	3495	0.76 (0.75–0.77)	0.93 (0.92–0.93)
Hispanic	71	0.71 (0.65–0.78)	0.43 (0.36–0.50)

lung cancer risk with an AUC of 0.88. It was unclear if the model predicted new cases, particularly when only predicting risk over a 1-year time frame. The only exclusion criterion was a past diagnosis of lung cancer, and predictors included lung cancer symptoms and diagnostic procedures. The other published model (7) based on Kaiser Permanente EHR data reported a high AUC of 0.86 but followed a case–control design which should be prospectively validated using a cohort design and recalibrated as a best practice (23). Both models included older patients, with age being a top predictor in both models.

The current study attempts to create a framework for EHR-based lung cancer prediction models. The cohort design, strict exclusion criteria, and operationalization of the outcome ensured that the model predicted risk of new cancer and not prevalent cancer, demonstrated by non-clustering of cases close to index. The validation of the model using two different datasets, including an EHR and a claims dataset, and its availability for use on an open-source platform are additional enhancements of this model compared with prior work.

In this study, the model failed to discriminate Asian and Hispanic lung cancer patients with similarly high probabilities when compared with non-Asian and non-Hispanic subgroups. To our knowledge, this is the first lung cancer prediction study to incorporate analyses examining racial and ethnic disparities in model performance using xAUCs, thus following recommendations to assess if health inequities are further perpetuated by machine learning models using historical data (24). Using xAUCs enabled investigation of potential systematic racial biases in model discrimination, which could reflect systematic under capture of data for certain race and ethnic groups and the observed lower incidence of lung cancer in Hispanics and Asians compared with non-Hispanic White individuals (9). It is possible that EMR-based machine learning models may ultimately need to be nested within each race and ethnicity subgroup, rather than on the overall patient population. Future efforts could involve calibrating models in datasets enriched with Asian and Hispanic patients.

A simple model that excluded race and only included age and smoking performed poorer than the EHR model but appeared to show slightly less bias across races and ethnicities. However, the efficiencies offered by a fully automated high-dimensional model that includes a broader health history, an incremental discriminatory performance, and a better model fit are important advantages of the full EHR-based model. Nevertheless, investigating ways to balance parsimony of covariates versus performance loss in discrimination and calibration will be a valuable area of future research.

Some of the most informative features of age, smoking, ethnicity, and COPD were also strong predictors in the two previously published EHR-based machine learning models (7, 8). As expected, smoking status was a very important predictor for lung cancer risk. EHR-based models have the advantage of capturing smoking status of a patient, rather than solely relying on administrative claims-based diagnosis codes. The higher AUC (0.81) when validating the model on Mercy EHR compared with the lower AUC (0.72) when validating on Optum claims dataset is likely suggestive of the smoking status variable being more informative than just diagnosis codes for tobacco dependence. However, healthcare organizations such as payers and health insurance companies that only have access to an administrative-claims dataset can continue to use this model to identify their high-risk enrollees for screening and obtain a reasonable discriminatory ability, despite limited smoking capture.

The automated lung cancer prediction modeling framework proposed in this study can identify a high-risk patient and trigger the physician to initiate a conversation with the patient. Institutions can select risk thresholds that align with their own goals of either narrowing the high-risk subset further or increasing sensitivity and broadening the pool for screening eligibility. Health care institutions are increasingly deploying automated clinical decision support systems using their EHR systems. Open access to analytic packages and software code disseminated in the current study will facilitate adoption in clinical practice and validation using an institution's own data as often as needed; especially where early lung cancer detection and diagnosis is a priority.

Strengths of this lung cancer risk prediction model include the large sample, validation on external databases and different database formats, temporal consistency in discriminatory ability, and the ability to identify a 'high-risk' subset even in a low incidence population. The high dimensionality of the model precluded subjective selection of covariates, while the open sharing of model and code enables further external validation.

The main limitation of this model is the lack of pathologic confirmation of lung cancer, necessitating a real-world data definition of lung cancer using diagnosis codes. Nevertheless, the lung cancer incidence (0.1%) in this study is consistent with the lung cancer incidence in this age group in the US population (9). The study datasets involved a disproportionately high distribution of individuals of White race and non-Hispanic ethnicity, similar to other large databases in the US. Finally, there may have been potential overlap of data used for the main model and the 2014 validation cohort, but maintenance of model performance for all later cohorts suggests the impact of data overlap to be minimal if any.

A recent editorial (25) highlighted the limitations regarding the retrospective nature of EHR-based models and the lack of randomized controlled trials showing evidence of efficacy of these models. Although the effectiveness of implementing this model in actual clinical practice is yet to be studied, there are potentially multiple applications of this prediction modeling framework to improve outcomes for lung cancer, by being an efficient complement to popular clinical risk prediction models.

Authors' Disclosures

U. Chandran reports receiving salary and compensation from Johnson & Johnson LLC; the submitted work was funded by Johnson & Johnson LLC. J. Reps reports being an employee of Janssen R&D and shareholder of Johnson & Johnson. R. Yang reports other support from Johnson & Johnson outside the submitted work. A. Vachani reports personal fees from Johnson & Johnson during the conduct of the study; grants from Precyte Inc., MagArray Inc., Optellum Ltd., Gordon and Betty Moore Foundation; and grants from Lungevity outside the submitted work. F. Maldonado received fees from Johnson & Johnson for preparation of this manuscript. I. Kalsekar reports other support from Johnson & Johnson outside the submitted work.

Authors' Contributions

U. Chandran: Conceptualization, supervision, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing. **J. Reps:** Resources, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–review and editing. **R. Yang:** Conceptualization, writing–review and editing. **A. Vachani:** Investigation, writing–review and editing. **F. Maldonado:** Investigation, writing–review and editing. **I. Kalsekar:** Conceptualization, resources, investigation, methodology, writing–review and editing.

Acknowledgments

This work was funded by Johnson & Johnson, LLC.

The authors would like to acknowledge Dr. Jay Carlson and Kerry Bommarito at Mercy Health Systems for their help with getting access to de-identified EHR data for the external validation cohort.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Note

Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Received August 12, 2022; revised October 7, 2022; accepted December 19, 2022; published first December 28, 2022.

References

1. U.S. Cancer Statistics Working Group. US Cancer Statistics Data Visualizations Tool, based on 2021 submission data (1999–2019): US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2022. Available from: www.cdc.gov/cancer/dataviz.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72:7–33.
3. Fedewa SA, Bandi P, Smith RA, Silvestri GA, Jemal A. Lung cancer screening rates during the COVID-19 pandemic. Chest 2022;161:586–9.
4. Wang Y, Midthun DE, Wampfler JA, Deng B, Stoddard SM, Zhang S, et al. Trends in the proportion of patients with lung cancer meeting screening criteria. JAMA 2015;313:853–5.
5. US Preventive Services Task Force. Clinician summary of USPSTF recommendation: screening for lung cancer 2021. Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening.
6. Faselis C, Nations JA, Morgan CJ, Antevil J, Roseman JM, Zhang S, et al. Assessment of lung cancer risk among smokers for whom annual screening is not recommended. JAMA Oncol 2022:e222952.
7. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine learning for early lung cancer identification using routine clinical and laboratory data. Am J Respir Crit Care Med 2021;204:445–53.
8. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, et al. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine. J Med Internet Res 2019;21:e13260.
9. National Cancer Institute Surveillance Epidemiology and End Results Program. Cancer Stat Facts: Lung and Bronchus Cancer: National Cancer Institute; 2022. Available from: https://seer.cancer.gov/statfacts/html/lungb.html.
10. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 2015;13:1.
11. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between Medicare claims and cancer registry data. Cancer Causes Control 2007;18:561–9.
12. Goldsbury D, Weber M, Yap S, Banks E, O'Connell DL, Canfell K. Identifying incident colorectal and lung cancer cases in health service utilization databases in Australia: a validation study. BMC Med Inform Decis Mak 2017;17:23.
13. Berquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data Proc Mach Learn Res 2017;68:25–38.
14. Turner RM, Chen YW, Fernandes AW. Validation of a case-finding algorithm for identifying patients with non–small cell lung cancer (NSCLC) in administrative claims databases. Front Pharmacol 2017;8:883.
15. Hardin J, Reps JM. Evaluating the impact of covariate lookback times on performance of patient-level prediction models. BMC Med Res Methodol 2021;21:180.
16. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018;25:969–75.
17. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol 2020;20:102.
18. Reps JM, Ryan P, Rijnbeek PR. Investigating the impact of development and internal validation design when training prognostic models using a retrospective cohort in big US observational healthcare data. BMJ Open 2021;11:e050146.
19. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.
20. Kallus N, Zhou A. The fairness of risk scores beyond classification: bipartite ranking and the xAUC metric. Available from: https://arxiv.org/pdf/1902.05826.pdf.
21. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernandez-Bertolin S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed 2021;211:106394.
22. Ten Haaf K, Bastani M, Cao P, Jeon J, Toumazis I, Han SS, et al. A comparative modeling analysis of risk-based lung cancer screening strategies. J Natl Cancer Inst 2020;112:466–79.
23. Reps JM, Ryan PB, Rijnbeek PR, Schumie MJ. Design matters in patient-level prediction: evaluation of a cohort vs. case–control design when developing predictive models in observational healthcare datasets. J Big Data 2021;8:1–18.
24. Rojas JC, Fahrenbach J, Makhni S, Cook SC, Williams JS, Umscheid CA, et al. Framework for integrating equity into machine learning models: a case study. Chest 2022;161:1621–7.
25. Pinsky P. Electronic health records and machine learning for early detection of lung cancer and other conditions: thinking about the path ahead. Am J Respir Crit Care Med 2021;204:389–90.