Can Historical Electronic Health Record Smoking Data Be Leveraged to Improve the Selection of Patients for Lung Cancer Screening?

Daniel J. Kats, BS[1,2]; Abdulhakim Tlimat, MD[2]; Yosra Adie, MPH[2,4]; Yasir Tarabichi, MD, MSCR[2,3]

[1]School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA; [2]Center for Clinical Informatics Research and Education, [3]Division of Pulmonary, Critical Care, and Sleep Medicine, MetroHealth Campus of Case Western Reserve University, Cleveland, Ohio, USA; [4]Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA

Introduction

Electronic health record (EHR) data provides healthcare systems an opportunity to leverage sophisticated prediction models to better select patients for lung cancer screening. However, smoking data in the EHR is notoriously inaccurate and the efficacy of the aforementioned models with real-world EHR data is unknown[1,2]. In this study, we had two goals:

1. To validate two leading models[3,4] using historical EHR data from an academic community healthcare system
2. To assess whether the development of new summary metrics could compensate for known EHR data limitations

We hypothesized that developing metrics that account for a patient's smoking *history*, as opposed to current smoking duration and intensity, would improve the predictive power of current risk models.

Methods

Study Design and Setting

This was a retrospective cohort study performed in an urban, safety-net medical system (including over a dozen ambulatory centers) in Cuyahoga county of northeast Ohio. We reviewed records of all outpatient encounters of every smoker between 1999 and 2018 (1,194,684 visits with 44,100 patients).

Statistical Analysis

For every outpatient visit of every smoker, we calculated "history-conscious" metrics:

- Average smoking intensity $= \dfrac{\text{total exposure}}{\text{total duration}} = \dfrac{\sum(\text{smoking intensity} \times \text{smoking duration})}{\sum(\text{smoking duration})}$
- Auto-incrementing duration: duration automatically increases by 1/365 every day without being manually updated

Lung cancer risk scores (United States Preventive Services Task Force (USPSTF) eligibility, $\text{PLCO}_{\text{M2012}}$[5], and LCRAT[6]) were calculated for each encounter with both raw EHR data and our history conscious metrics. The incidence of lung cancer was detected by diagnostic coding and confirmed for accuracy by encounter narrative review.

Analysis was performed using the RStudio software package version 1.0.143 (RStudio Team, Boston, MA) and R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

Population

Characteristic	No Cancer No.	No Cancer %	Cancer No.	Cancer %
n	42,841	97.1	1,259	2.9
Sex				
Female	22,095	51.6	688	54.6
Male	20,746	48.4	571	45.4
Race/ethnicity				
White	22,558	52.7	707	56.2
Black	16,202	37.8	495	39.3
Hispanic	1,980	4.6	29	2.3
Other	388	0.9	10	0.8
Education				
< HS	2,218	25.5	77	34.4
HS/GED	3,386	38.9	91	40.6
Some training after HS	20	0.2	1	0.4
Some college (< 4y)	1,983	22.8	38	17.0
Bachelor's degree	753	8.6	13	5.8
> Bachelor's degree	353	4.1	4	1.8
Insurance				
Commercial	10,325	24.5	184	14.9
Medicare	14,430	34.2	631	51.0
Medicaid	11,863	28.1	282	22.8
Other	5,583	13.2	140	11.3

Table 1 – Demographics of study patients. Other demographics (e.g. age, smoking data, etc.) were not included in this table because they can fluctuate over different encounters for the same patient.

Results

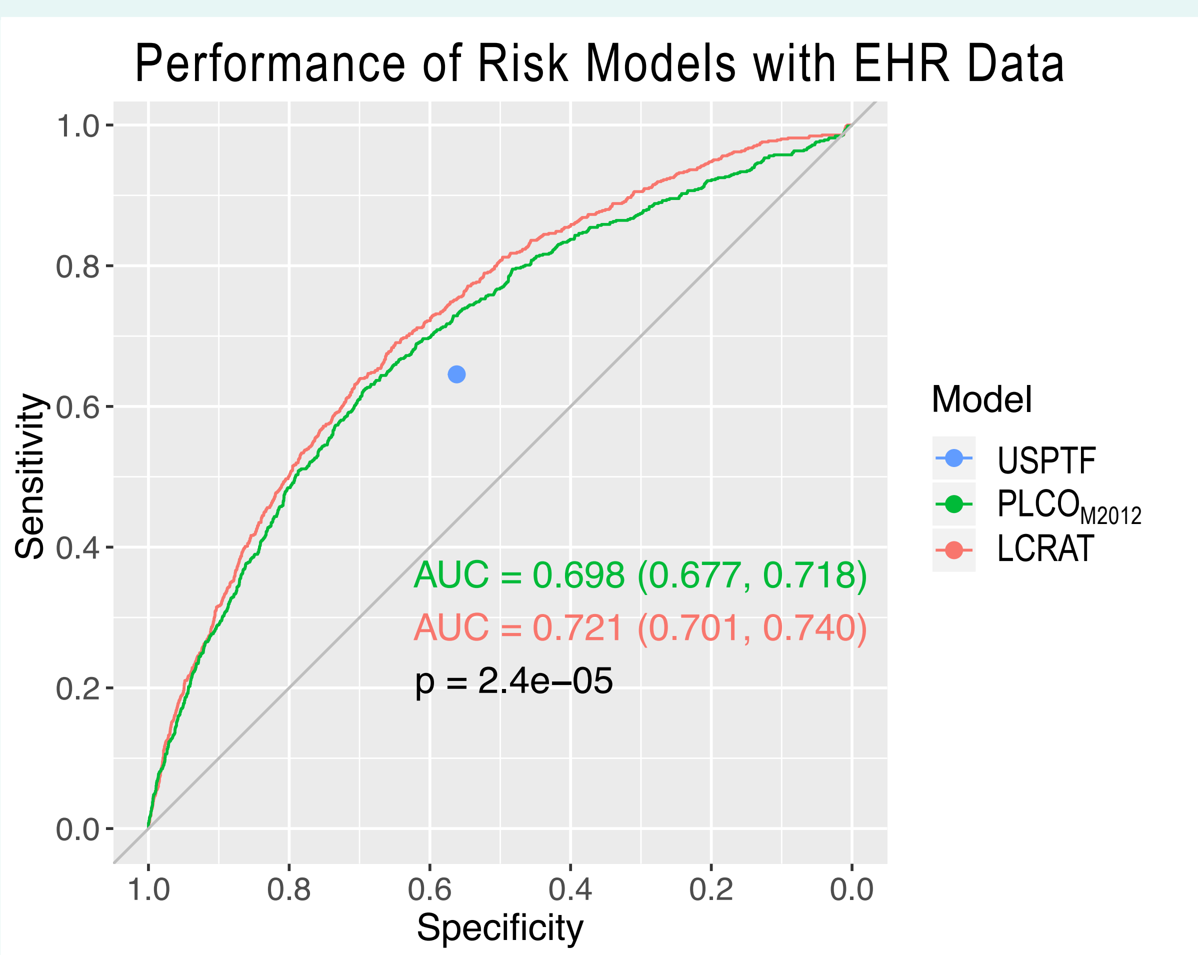

Figure 1 – Receiver operator characteristic (ROC) curves were computed for the two models assessed in our population, with the sensitivity and specificity of USPSTF guidelines superimposed for comparison. The above scores were calculated with raw EHR data. The curves and areas under the curve were similar when scores were calculated with our history-conscious metrics (not shown).

Both $\text{PLCO}_{\text{M2012}}$ and LCRAT performed similarly reasonably (according to area under the curve), but outperformed the national USPSTF screening guidelines (Figure 1). However, using history-conscious metrics did *not* improve the results, contrary to our expectations ($\text{PLCO}_{\text{M2012}}$ ROC curves are shown in Figure 2; LCRAT curves are not shown, but behave similarly).

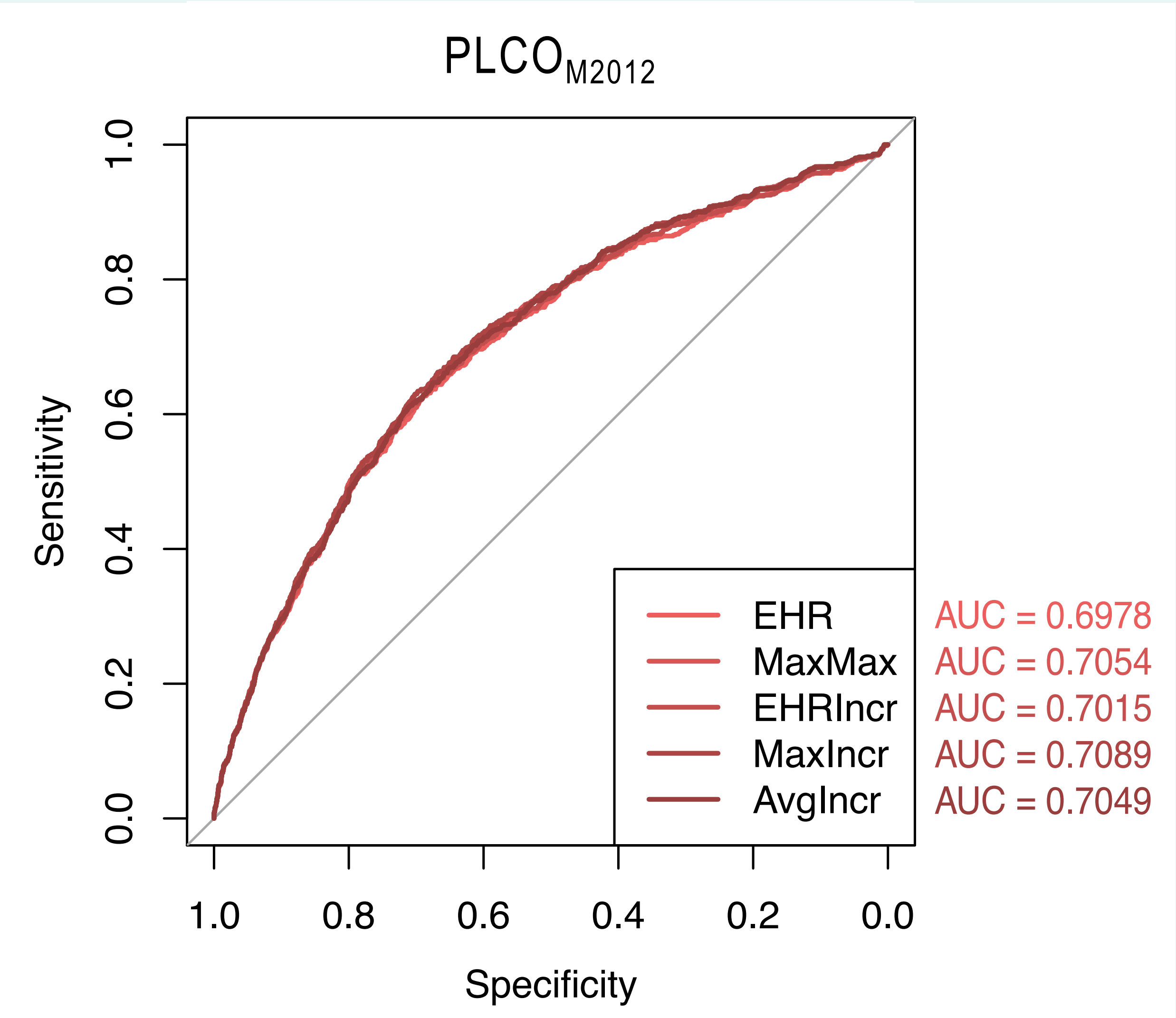

Figure 2 – ROC curves were computed with various combinations of EHR and history-conscious smoking metrics. "EHR" denotes raw smoking intensity and duration; "MaxMax" denotes lifetime maximum smoking intensity and duration; "EHRMax" denotes raw smoking intensity and lifetime maximum duration; "MaxIncr" denotes lifetime maximum smoking intensity and auto-incrementing duration; "AvgIncr" denotes average smoking intensity and auto-incrementing duration.

Using a $\text{PLCO}_{\text{M2012}}$ lung cancer risk threshold of 1.5% (with respect to the chance of developing lung cancer within six years) would recommend 5,723 out of 11,576 USPSTF-eligible smokers in our study population for screening, potentially detecting lung cancer in 537 subjects who would otherwise go on to manifest the diagnosis within a six-year period. When substituting our history-conscious metrics for standard EHR data and calibrating the risk threshold to identify the same number of patients for screening, only 5,500 patients (3.9% fewer) would need to be screened. On the other hand, the performance of the LCRAT model was not affected by substitution of history-conscious metrics for EHR data.

Discussion

Despite known issues in the accuracy of patient demographic and smoking history data in the EHR, the $\text{PLCO}_{\text{M2012}}$ and LCRAT models performed reasonably well in our real-world EHR-based validation. While a history-conscious approach should logically improve the performance of prediction models, the addition of our history-conscious metrics led to a non-clinically significant improvement in discrimination. Our metrics also failed to improve the AUCs of the models ROC curves. Incorporating lung cancer risk calculators into the EHR (either with or without a history-conscious approach) could lead to more efficient screening and less false positive test burden compared to USPSTF guideline recommendations. Prospective validation is necessary to determine the practical and effective utility of such an approach.

Future Directions

- Will a model trained with EHR data outperform published models that were trained with gold standard survey data?
- Are there other history conscious metrics that can augment current models?
    - Smoking trend (increasing vs. decreasing intensity)
- Will primary care providers use risk calculators built into the EHR?
    - Will this result in fewer false positives without fewer true positives?

References

1) Kinsinger, L.S., Anderson, C., Kim, J., Larson, M., Chan, S.H., King, H.A., Rice, K.L., Slatore, C.G., Tanner, N.T., Pittman, K., Monte, R.J., McNeil, R.B., Grubber, J.M., Kelley, M.J., Provenzale, D., Datta, S.K., Sperber, N.S., Barnes, L.K., Abbott, D.H., Sims, K.J., Whitley, R.L., Wu, R.R., Jackson, G.L. Implementation of Lung Cancer Screening in the Veterans Health Administration. *JAMA Intern Med*. 2017, 177(3): 399-406.
2) Modin, H.E., Fathi, J.T., Gilbert, C.R., Wilshire, C.L., Wilson, A.K., Aye, R.W., Farivar, A.S., Louie, B.E., Vallieres, E., Gorden, J.A. Pack-Year Cigarette Smoking History for Determination of Lung Cancer Screening Eligibility: Comparison of the Electronic Medical Record versus a Shared Decision Making Conversation. *Ann Am Thorac Soc*. 2017.
3) Katki, H.A., Kovalchik, S.A., Petito, L.C., Cheung, L.C., Jacobs, E., Jemal, A., Berg, C.D., Chaturvedi, A.K. Implications of Nine Risk Prediction Models for Selecting Ever-Smokers for Computed Tomography Lung Cancer Screening. *Ann Intern Med*. 2018.
4) Ten Haaf, K., Jeon, J., Tammemägi, M.C., Han, S.S., Kong, C.Y., Plevritis, S.K., Feuer, E.J., de Koning, H.J., Steyerberg, E.W., Meza, R. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. *PLoS Med*. 2017, 14(4): e1002277.
5) Tammemägi, M.C., Katki, H.A., Hocking, W.G., Church, T.R., Caporaso, N., Kvale, P.A., Chaturvedi, A.K., Silvestri, G.A., Riley, T.L., Commins, J., Berg, C.D. Selection criteria for lung-cancer screening. *N Engl J Med*. 2013, 368(8): 728-736.
6) Katki, H.A., Kovalchik, S.A., Berg, C.D., Cheung, L.C., Chaturvedi, A.K. Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. *JAMA*. 2016, 315(21): 2300-2311.