Using Health Administrative Data to Predict Chronic Obstructive Pulmonary Disease Exacerbations

Lili Jiang, M.D., M.P.H., M.Sc., Ph.D.[1,2,3], and Andrea S. Gershon, M.D., M.Sc.[1,2,3]

[1]Sunnybrook Research Institute, Toronto, Ontario, Canada; [2]ICES, Toronto, Ontario, Canada; and [3]Department of Medicine, University of Toronto, Toronto, Ontario, Canada

ORCID ID: 0000-0002-0246-594X (A.S.G.).

Chronic obstructive pulmonary disease (COPD) exacerbations are key events in the course of COPD that place both health and economic burdens on individuals and society (1). Exacerbations not only accelerate decline in lung function, cause disease progression, impair quality of life, and increase mortality, but they are also the main drivers of healthcare use such as emergency department visits and hospitalizations (2). Hence, exacerbation prevention is a key goal of COPD management. Recent efforts have been made to develop predictive models to identify patients at high risk for COPD exacerbations because prediction could not only help prevent COPD exacerbations but also facilitate early treatment, shorten the length of exacerbations, and lessen exacerbation impact. According to a systematic review of predictions models for COPD exacerbations, most existing models are based on clinical point-of care data (3). The application of data collected for administrative or billing purposes that widely cover areas like demographic information, drugs, physician services, and hospital services, in these types of prediction models is limited. As computing technology and predictive modeling advance, it becomes possible to use the immense volume of healthcare administrative data in healthcare research and surveillance, including in real-time predictions to alert people of health risks.

In this issue of *AnnalsATS*, Tavakoli and colleagues (pp. 1069–1076) conducted a proof-of-concept study that developed models to predict patients at high-risk for hospitalization for acute exacerbations of COPD (4). The study used healthcare administrative data from between 1997 and 2016 from the province of British Columbia in Canada, which has a universal, provincially funded medical insurance plan that covers most healthcare and pharmacy costs for the entire population. The authors aimed to use data from a 6-month time period to predict the risk of severe COPD exacerbation in the subsequent 2 months. Prediction models were developed based on a cohort of 108,433 patients and validated by an internal temporal cohort (the same source data as the development cohort but at a later time period) of 113,786 patients. A total of 1,126 and 1,136 people from these two cohorts, respectively, were hospitalized because of COPD within their outcome windows. By using traditional statistical methods and machine learning algorithms, the investigators found health administrative models that had better predictive abilities than a reference model that only included previous exacerbations history in terms of both discrimination and calibration. The best prediction model (gradient boosting), for example, had an area under the receiver operating characteristic curve of 0.82 (95% confidence interval 0.80–0.83), whereas the model with exacerbation history as the only predictor had an area under the receiver operating characteristic curve of 0.68 (95% confidence interval 0.67–0.69). A method called least absolute shrinkage and selection operator was used to select variables, using an approach that avoided including too many variables in the models or overfitting.

Tavakoli and colleagues have confirmed the potential application of healthcare administrative data in public health surveillance of COPD exacerbations. Healthcare administrative datasets have been used in the past to determine risk factors for COPD exacerbation (5, 6). However, these common casual analyses examine the association between potential risk factors and a health event on a “population level” but do not quantify an individual’s risk of a health outcome given their characteristics. Predictive algorithms allow us to use such characteristics to identify or measure an individual’s probability of experiencing a health outcome (7). Hence, physicians, public health professionals, and policy makers could use the COPD exacerbation predictive models developed to identify high-risk individuals for specific disease management through phone consultation, home visits, and other support to prevent or delay exacerbations. In addition, the predicted number of patients at risk could also help public health authorities determine their ability to manage the problem. In turn, at-risk individuals would be made aware of their situation, be able to institute their own self-management processes (if they had them), and feel reassured by the help.

Some types of health administrative data have many strengths in the context of COPD exacerbation prediction. They cover

DOI: 10.1513/AnnalsATS.202006-704ED

Ann Am Thorac Soc Vol 17, No 9, pp 1056–1061, Jul 2020
Internet address: www.atsjournals.org

a majority of the population and are less likely to have missing values (8). They are continuously updated over time, which allows the predicted risk to also be updated over time and accurate forecasts to be provided (9). Clinical point-of-care data contains more clinical detail but have other problems. Missing information and variations among different study sites cause problems for clinical data integrity and interpretation. Clinical data also often target groups of select patients and are collect at a specific time point.

Using healthcare administrative data to predict COPD exacerbations also has some limitations. First, timeliness of data availability affects how useful it is for prediction. There is always a lag time between when a healthcare encounter occurs and when information about it is available for analysis (10). For example, the availability of data related to healthcare system use tends to follow billing cycles. Vital statistics data become available only after data are regularly processed (entered and coded). However, this lack of timeliness affects real-time prediction. Tavakoli and colleagues aimed to use data in a 6 months period to predict COPD exacerbations in the following 2 months. This would be problematic if the dataset only became available 3 months after a healthcare encounter. Second, the accuracy and reliability of prediction models largely depend on available predictors. Although studies have shown that models based on healthcare administrative data were comparable with those derived from clinical databases, clinical specificity for conditions and laboratory results are more useful in predicting short-term healthcare outcomes (11). Developing standardized electronic medical records and including them in predictive model development must be a future direction in this area. Finally, although machine learning has been increasingly used in public health research and has yielded some impressive practical successes (12), its application is limited in its ability to "explain" its predictions in an understandable way (13). Although the underlying mathematical principles of such models are understandable, it is difficult and often impossible to interrogate the inner workings of models to follow how and why they made a certain prediction. This is especially problematic for clinical professionals who have particular demand for approaches that are not only well performing but also biologically interpretable and explainable.

In sum, the advance of computer science and health administrative data analysis has potential to be used to forecast the real-time adverse health outcomes of chronic diseases, such as COPD exacerbations. However, the accuracy of the prediction is dependent on the timeliness of data and would be enriched if it included clinical predictors. More efforts are required to explain how this approach works to obtain acceptance by clinicians. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1 World Health Organization. Chronic obstructive pulmonary disease (COPD) [Internet]. 2017 [accessed 2020 Jul 17]. Available from: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

2 Halpin DM, Miravitlles M, Metzdorf N, Celli B. Impact and prevention of severe exacerbations of COPD: a review of the evidence. *Int J Chron Obstruct Pulmon Dis* 2017;12:2891–2908.

3 Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. *Eur Respir Rev* 2017;26:160061.

4 Tavakoli H, Chen W, Sin DD, FitzGerald JM, Sadatsafavi M; Canadian Respiratory Research Network. Predicting severe chronic obstructive pulmonary disease exacerbations: developing a population surveillance approach with administrative data. *Ann Am Thorac Soc* 2020;17:1069–1076.

5 Stanford RH, Nag A, Mapel DW, Lee TA, Rosiello R, Vekeman F, *et al*. Validation of a new risk measure for chronic obstructive pulmonary disease exacerbation using health insurance claims data. *Ann Am Thorac Soc* 2016;13:1067–1075.

6 Stanford RH, Nag A, Mapel DW, Lee TA, Rosiello R, Schatz M, *et al*. Claims-based risk model for first severe COPD exacerbation. *Am J Manag Care* 2018;24:e45–e53.

7 Ranapurwala SI, Cavanaugh JE, Young T, Wu H, Peek-Asa C, Ramirez MR. Public health application of predictive modeling: an example from farm vehicle crashes. *Inj Epidemiol* 2019;6:31.

8 Cadarette SM, Wong L. An introduction to health care administrative data. *Can J Hosp Pharm* 2015;68:232–237.

9 Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2: 230–243.

10 Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data Sci J* 2015;14:2.

11 Zeltzer D, Balicer RD, Shir T, Flaks-Manov N, Einav L, Shadmi E. Prediction accuracy with electronic medical records versus administrative claims. Stanford, CA: Stanford University; 2019 [accessed 2020 May 25]. Available from: https://web.stanford.edu/~leinav/pubs/MC2019.pdf.

12 Sanchez-Morillo D, Fernandez-Granero MA, Leon-Jimenez A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. *Chron Respir Dis* 2016;13:264–283.

13 Medicine TLR; The Lancet Respiratory Medicine. Opening the black box of machine learning. *Lancet Respir Med* 2018;6:801.