Machine Learning Applied to Routinely Collected Health Administrative Data

Laura C. Rosella and Vinyas Harish

Abstract

There has been considerable growth in the development of machine learning algorithms for clinical applications. The authors survey recent machine learning models developed with the use of large health administrative databases at ICES and highlight three areas of ongoing development that are particularly important for health system applications.

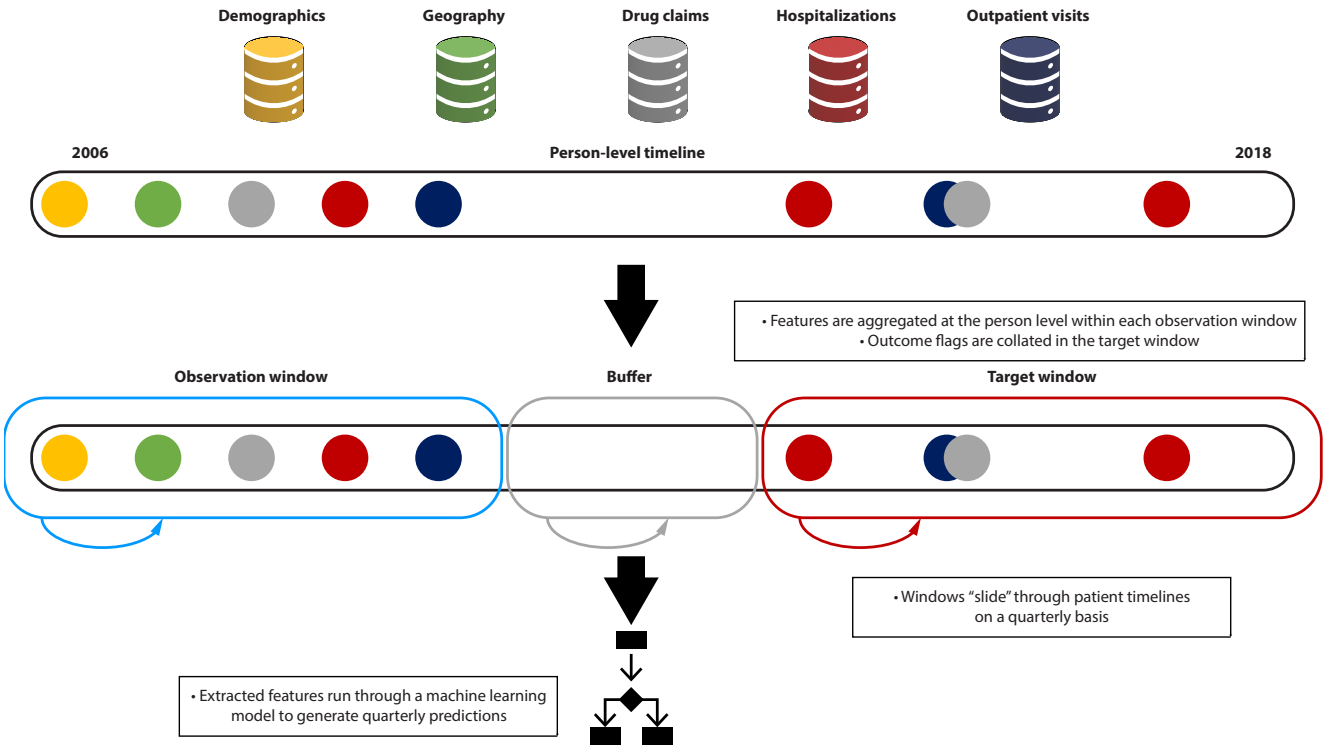
Introduction

There has been increasing interest in applying machine learning methods to large, linked and routinely collected health administrative databases that contain person-level data on billing codes, procedures, medications, geography and demographics. Machine learning models have been gaining in popularity as there is increasing availability of large databases that support methods that require fewer assumptions in underlying data structure and less user input in the selection of predictors, and focus on maximizing predictive performance as opposed to supporting hypothesis-driven inference (Bi et al. 2019). Most recent health database examples are focused on a supervised classification task (i.e., generating a probability of a patient experiencing a disease state or healthcare event), but there also have been examples of unsupervised applications (i.e., to reveal subgroups of significance for population segmentation or to understand patient heterogeneity) (Liao et al. 2016; Morgenstern et al. 2020; Schäfer et al. 2010). What insights are possible with applying these more flexible methods to large health administrative databases? How can these models inform policy and clinical practice? And finally, what are the key considerations when developing and using these models? We address these questions through a survey of recent models developed using ICES data holdings in Ontario (Gutierrez et al. 2021; Ravaut et al. 2021a, 2021b; Yi et al. 2021).

Increasing Application of Machine Learning Methods to Routinely Collected Health Data

Despite enthusiasm for machine learning methods, there has been considerable debate on where value is added compared to traditional statistical approaches (Christodoulou et al. 2019; Weaver and McAlister 2021). One of the challenges with these comparisons is that they are focused on a heterogeneous set of models with wide variability in the properties of the data. Machine learning approaches are more likely to show an advantage when applied to a more complex data structure where the flexibility models can demonstrate value. Ravaut et al. (2021b) demonstrated one approach to leveraging this complexity by gathering healthcare events across multiple health administrative data sets chronologically and mapping them dynamically to a patient timeline (Ravaut et al. 2021b) (Figure 1). A high level of model performance was achieved by learning the relationship between a patient's health state and their risk of an outcome – in this case, developing complications from diabetes. This learning has to be undertaken dynamically as each individual's interactions with the health system varies over time. This "sliding window" approach allowed for data from millions of patients and tens of millions of patient instances to train the model, ultimately resulting in its ability to accurately predict the risk of a patient developing complications from diabetes three years in advance with high performance.

FIGURE 1.
Structure of health administrative data used in the development of the machine learning model



Source: Adapted from Ravaut et al. 2021b.

Potential for Health Services and Policy

There has been considerable growth in the development of machine learning algorithms for individual clinical applications; however, less attention has been paid to system-level applications. Largely due to social determinants, the risks of numerous adverse health outcomes (e.g., the development of chronic diseases) are not diffuse in a population but concentrated in certain subgroups. Data-driven tools can be used to support health systems for ongoing risk assessment and the targeting of interventions for those who would benefit the most (Manuel and Rosella 2010; Rose 2020). Further applications include understanding cost distributions to improve system management (Bilandzic and Rosella 2017; O'Neill et al. 2021; Ravaut et al. 2021a).

One example of system-level applications was demonstrated in the context of the COVID-19 pandemic. Upon the authorization of COVID-19 vaccines in early 2021, several health systems needed to determine how to prioritize the vaccine rollout. To prioritize vaccine eligibility, many jurisdictions developed simple decision rules involving age, geography and co-morbidity burden. We developed a machine learning model using two years' worth of data on demographic characteristics, prescriptions, laboratory values and health system interactions to predict the risk of hospitalization in patients who tested positive for SARS-CoV-2 (Gutierrez et al. 2021). The performance of this model was tested against four empirical rules: ranking by age, ranking by the number of comorbidities and two sequential combinations of age and comorbidities using recall (or sensitivity) at three cut-off points (the top 10%, 20% and 30% of patients, generally referred to as "recall at k"). The analysis includes a comparison of model-based recall versus empirical rules to inform vaccine prioritization based on age or comorbidity and thus demonstrated that the additional information included in the machine learning model could allow for more refined decision making at the population level.

This approach holds promise for many health system applications. Patients with chronic diseases drive the majority of healthcare spending (Tinetti et al. 2012), yet our healthcare and public health systems often take siloed approaches to chronic disease management and prevention (Rosella and Kornas 2018). For example, patients with diabetes are at risk for a wide range of complications including cardiovascular events, kidney disease and retinopathy. Ravaut et al. (2021b) combined the aforementioned "recall at k" metric of their developed machine learning algorithm with a validated costing

algorithm to estimate the cost burden of patients developing a range of diabetes-related complications. The top 1% of patients predicted at risk by this model represents a cost of over $400 million. Targeting interventions based on such model outputs could help maximize the cost-effectiveness of limited health system resources and meaningfully impact disease trajectories. Overall, these approaches support the potential for more precise stratification of risk groups in the population. These insights must be balanced with a population-wide perspective that focuses on policies and interventions that address system and societal factors, in addition to individual-level clinical interventions (Ramaswami et al. 2018).

Methodological Considerations and the Need for Teamwork

One common thread in the successful implementation of machine learning models as they apply to health data is the need for multidisciplinary teams that include epidemiologists, statisticians, computer scientists, clinicians and other experts who deeply understand the nuances of the data. In addition, successful projects will have both a health system perspective – especially if that is the intended application – and a clinical perspective as many of the data are generated from clinical encounters (Agniel et al. 2018). This mix is essential not only to ensure both rigorous development and analysis but also to create relevant and impactful outcomes. Supporting these multidisciplinary teams might also require building capacity in modern computing infrastructure (Schull et al. 2020).

Numerous metrics are used to measure the performance of predictive models. In clinical settings, two commonly used metrics are precision (or positive predictive value) and recall. There are often trade-offs in optimizing performance with these metrics because, in clinical settings, stakeholders need to balance the detection of enough patients at risk for certain outcomes with alarm fatigue (Verma et al. 2021). However, at the health systems level, the concordance between observed and predicted risks, otherwise known as calibration, is particularly important (Van Calster et al. 2019). Underestimating risk with a predictive model meant to guide resource allocation could amplify gaps in care that may already exist, while overestimating risk could lead to inefficient use of resources and not adequately addressing need in the population (Chen et al. 2020).

A major issue with the application of risk models in population health is their ability to recapitulate biases in underlying data and amplify health inequities (Gervasi et al. 2022). These biases, which have been noted in existing algorithms in the US, perpetuate inequities in who is receiving care and covered through health insurance (Obermeyer et al. 2019). A number of strategies to minimize the likelihood of bias in health algorithms are described elsewhere (Chen et al. 2021), but one important approach involves examining models for performance gaps in population subgroups to ensure that the model is performing similarly across the population. In our work to predict the onset of type 2 diabetes (Ravaut et al. 2021a), we examined calibration not just in the entire test population but also by age, material deprivation, race/ethnicity, sex, amount of contact with the healthcare system and immigration status, demonstrating subgroup performance of a diabetes-onset prediction model across population groups. The model is evaluated on several partitions of the test population and for each subset reports the size, incidence rate and average model prediction. This type of analysis is essential in model assessment because if discrepancies are found in subgroup performance then the decisions based on these algorithms will differentially affect subgroups in the population. When these differences are noted, there may be a need to revisit model development or apply the use of recalibration methods to correct these discrepancies and increase fairness (Barda et al. 2020).

Continued Challenges to Overcome

As the applications for machine learning increase, there will be continual improvements to the underlying methods. We identify three areas of ongoing development that are particularly important for health system applications. Interpretability (understanding how the model parameters are influencing the model outcome) and explainability (accounting for how the algorithm influences decisions made based on that model) are two complementary approaches to foster trust in machine learning models (Petch et al. 2022). A third approach, transparency, documents how the model was created and validated to ensure that interpretability and explainability can be critically assessed, just as we would assess any evidence we use to inform health decisions (Andaur Navarro et al. 2021; Collins and Moons 2019). Tree-based models, such as the gradient-boosted decision trees we applied in our recent experience, work well for developing algorithms using administrative data sets, offer the advantage of being able to more easily report on model features and how they are used and are well suited to the underlying tabular structure of administrative data (Friedman 2002). Interpretability and explainability are areas of machine learning research that are rapidly evolving, especially through advances in visualization approaches (Apley and Zhu 2020; Lundberg et al. 2020; Vellido 2020).

The second challenge is related to the translation of these models. Once models are developed, health system practitioners must consider whether their end goal is to develop a static research product (i.e., a proof of concept) or an actively used tool to guide decision making. The latter goal requires insights into machine learning operations (MLOps) and human factors. MLOps are a set of practices to deploy and

maintain machine learning models in the real world (Alla and Adari 2021). A particular aspect of MLOps to consider is data set shift, or changes in the underlying data between when a model was developed and when it was used (Guo et al. 2021). For example, Gutierrez et al. (2021) developed their model prior to the emergence and spread of COVID-19 variants in Ontario. Because these variants have distinct biological, epidemiological and clinical profiles, the model's performance must be scrutinized over time to ensure it remains robust. Successfully deploying machine learning tools also requires thoughtful attention to several dimensions of human factors including user interfaces, user experiences and workflows (Sendak et al. 2020; Verma et al. 2021).

Finally, developers and users must recognize the limitations of health administrative data for certain applications. A clinical application, such as recognizing acute deterioration in hospitalized patients, necessitates using more granular and temporally resolved data from electronic medical records. Ongoing work to link administrative data to hospital, primary care and environmental databases will provide further insights into population health. Those working in health policy and health system settings must subject the data and the model outputs to careful scrutiny to determine if they will support their intended purpose.

Conclusion

As we begin to see more machine learning models generated from large health administrative databases, the questions we raise will become more important. Thoughtful practices in both the development and application of the models can support novel health system applications. We identify three key learnings needed for the continued evolution of machine learning methods to health administrative data. First is the need for explicit articulation of the intended use of the model that is appropriate given the nature of the data (i.e., coded data on healthcare interactions). This also includes clarifying how the generated model is appropriate for the intended purpose (e.g., health system applications versus informing direct patient care). Second, practitioners should adhere to methodological best practices to ensure rigour and transparency, including following reporting guidelines and risk-of-bias guidelines (Collins et al. 2021). In addition, models should be assessed through the lens of fairness and include thorough calibration and subgroup performance assessments. Finally, multidisciplinary teams are vital to ensure the successful development and assessment of these models, given the different perspectives needed on the methods, the content of the data and the intended applications in the health system. HQ

References

Agniel, D., I.S. Kohane and G.M. Weber. 2018. Biases in Electronic Health Record Data due to Processes within the Healthcare System: Retrospective Observational Study. *BMJ* 361: k1479. doi:10.1136/bmj.k1479.

Alla S. and S.K. Adari. 2021. *Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure*. Apress.

Andaur Navarro, C.L., J.A.A. Damen, T. Takada, S.W.J. Nijman, P. Dhiman, J. Ma et al. 2021. Risk of Bias in Studies on Prediction Models Developed Using Supervised Machine Learning Techniques: Systematic Review. *BMJ* 375: n2281. doi:10.1136/bmj.n2281.

Apley, D.W. and J. Zhu. 2020. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society: Series B* 82(4): 1059–86. doi:10.1111/rssb.12377.

Barda, N., G. Yona, G.N. Rothblum, P. Greenland, M. Leibowitz, R. Balicer et al. 2020. Addressing Bias in Prediction Models by Improving Subpopulation Calibration. *Journal of the American Medical Informatics Association* 28(3): 549–58. doi:10.1093/jamia/ocaa283.

Bi, Q., K.E. Goodman, J. Kaminsky and J. Lessler. 2019. What Is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology* 188(12): 2222–39. doi:10.1093/aje/kwz189.

Bilandzic, A. and L. Rosella. 2017. The Cost of Diabetes in Canada over 10 Years: Applying Attributable Health Care Costs to a Diabetes Incidence Prediction Model. *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice* 37(2): 49–53. doi:10.24095/hpcdp.37.2.03.

Chen, I.Y., S. Joshi and M. Ghassemi. 2020. Treating Health Disparities with Artificial Intelligence. *Nature Medicine* 26(1): 16–17. doi:10.1038/s41591-019-0649-2.

Chen, I.Y., E. Pierson, S. Rose, S. Joshi, K. Ferryman and M. Ghassemi. 2021. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4(1): 123–44. doi:10.1146/annurev-biodatasci-092820-114757.

Christodoulou, E., J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel and B. Van Calster. 2019. A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology* 110: 12–22. doi:10.1016/j.jclinepi.2019.02.004.

Collins, G.S., P. Dhiman, C.L.A. Navarro, J. Ma, L. Hooft, J.B. Reitsma et al. 2021. Protocol for Development of a Reporting Guideline (TRIPOD-AI) and Risk Of Bias Tool (PROBAST-AI) for Diagnostic and Prognostic Prediction Model Studies Based on Artificial Intelligence. *BMJ Open* 11: e048008. doi: 10.1136/bmjopen-2020-048008.

Collins, G.S. and K.G.M. Moons. 2019. Reporting of Artificial Intelligence Prediction Models. *The Lancet* 393(10181): 1577–79. doi:10.1016/s0140-6736(19)30037-6.

Friedman, J.H. 2002. Stochastic Gradient Boosting. *Computational Statistics and Data Analysis* 38(4): 367–78. doi:10.1016/s0167-9473(01)00065-2.

Gervasi, S.S., I.Y. Chen, A. Smith-McLallen, D. Sontag, Z. Obermeyer, M. Vennera et al. 2022. The Potential for Bias in Machine Learning and Opportunities for Health Insurers to Address It. *Health Affairs* 41(2): 212–18. doi:10.1377/hlthaff.2021.01287.

Guo, L.L., S.R. Pfohl, J. Fries, J. Posada, S. Lanyon Fleming, C. Aftandilian et al. 2021. Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine. *Applied Clinical Informatics* 12(4): 808–15. doi:10.1055/s-0041-1735184.

Gutierrez, J.M., M. Volkovs, T. Poutanen, T. Watson and L.C. Rosella. 2021. Risk Stratification for COVID-19 Hospitalization: A Multivariable Model Based on Gradient-Boosting Decision Trees. *CMAJ Open* 9(4): E1223–31. doi:10.9778/cmajo.20210036.

Liao, M., Y. Li, F. Kianifard, E. Obi and S. Arcona. 2016. Cluster Analysis and Its Application to Healthcare Claims Data: A Study of End-Stage Renal Disease Patients Who Initiated Hemodialysis. *BMC Nephrology* 17(1): 25. doi:10.1186/s12882-016-0238-2.

Lundberg, S.M., G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair et al. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* 2(1): 56–67. doi:10.1038/s42256-019-0138-9.

Manuel, D.G. and L.C. Rosella. 2010. Assessing Population (Baseline) Risk Is a Cornerstone of Population Health Planning—Looking Forward to Address New Challenges. *International Journal of Epidemiology* 39(2): 380–82. doi:10.1093/ije/dyp373.

Morgenstern, J.D., E. Buajitti, M. O'Neill, T. Piggott, V. Goel, D. Fridman et al. 2020. Predicting Population Health with Machine Learning: A Scoping Review. *BMJ Open* 10(10): e037860. doi:10.1136/bmjopen-2020-037860.

O'Neill, M., K. Kornas, W.P. Wodchis and L.C. Rosella. 2021. Estimating Population Benefits of Prevention Approaches Using a Risk Tool: High Resource Users in Ontario, Canada. *Healthcare Policy* 16(3): 51–66. doi:10.12927/hcpol.2021.26433.

Obermeyer, Z., B. Powers, C. Vogeli and S. Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366(6464): 447–53. doi:10.1126/science.aax2342.

Petch, J., S. Di and W. Nelson. 2022. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Canadian Journal of Cardiology* 38(2): 204–13. doi:10.1016/j.cjca.2021.09.004.

Ramaswami, R., R. Bayer and S. Galea. 2018. Precision Medicine from a Public Health Perspective. *Annual Review of Public Health* 39(1): 153–68. doi:10.1146/annurev-publhealth-040617-014158.

Ravaut, M., V. Harish, H. Sadeghi, K.K. Leung, M. Volkovs, K. Kornas et al. 2021a. Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. *JAMA Network Open* 4(5): e2111315. doi:10.1001/jamanetworkopen.2021.11315.

Ravaut, M., H. Sadeghi, K.K. Leung, M. Volkovs, K. Kornas, V. Harish et al. 2021b. Predicting Adverse Outcomes due to Diabetes Complications with Machine Learning Using Administrative Health Data. *NPJ Digital Medicine* 4(1): 24. doi:10.1038/s41746-021-00394-8.

Rose, S. 2020. Intersections of Machine Learning and Epidemiological Methods for Health Services Research. *International Journal of Epidemiology* 49(6): 1763–70. doi:10.1093/ije/dyaa035.

Rosella, L. and K. Kornas. 2018. Putting a Population Health Lens to Multimorbidity in Ontario. *Healthcare Quarterly* 21(3): 8–11. doi:10.12927/hcq.2018.25709.

Schäfer, I., E.C. von Leitner, G. Schön, D. Koller, H. Hansen, T. Kolonko et al. 2010. Multimorbidity Patterns in the Elderly: A New Approach of Disease Clustering Identifies Complex Interrelations between Chronic Conditions. *PLoS One* 5(12): e15941. doi:10.1371/journal.pone.0015941.

Schull, M., M. Brudno, M. Ghassemi, G. Gibson, A. Goldenberg, A. Paprica et al. 2020. Building a Research Partnership between Computer Scientists and Health Service Researchers for Access and Analysis of Population-Level Health Datasets. *International Journal of Population Data Science* 5(5): 1529. doi:10.23889/ijpds.v5i5.1529.

Sendak, M.P., M. Gao, N. Brajer and S. Balau. 2020. Presenting Machine Learning Model Information to Clinical End Users with Model Facts Labels. *NPJ Digital Medicine* 3(1): 41. doi:10.1038/s41746-020-0253-3.

Tinetti, M.E., T.R. Fried and C.M. Boyd. 2012. Designing Health Care for the Most Common Chronic Condition—Multimorbidity. *JAMA* 307(23): 2493–94. doi:10.1001/jama.2012.5265.

Van Calster, B., D.J. McLernon, M. van Smeden, L. Wynants and E.W. Steyerberg on behalf of Topic Group 'Evaluating Diagnostic Tests and Prediction Models' of the STRATOS Initiative. 2019. Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine* 17(1): 230. doi:10.1186/s12916-019-1466-7.

Vellido, A. 2020. The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care. *Neural Computing and Applications* 32(24): 18069–83. doi:10.1007/s00521-019-04051-w.

Verma, A.A., J. Murray, R. Greiner, J.P. Cohen, K.G. Shojania, M. Ghassemi et al. 2021. Implementing Machine Learning in Medicine. *CMAJ* 193(34): E1351–57. doi:10.1503/cmaj.202434.

Weaver, C.G.W. and F.A. McAlister. 2021. Machine Learning, Predictive Analytics, and the Emperor's New Clothes: Why Artificial Intelligence Has Not Yet Replaced Conventional Approaches. *Canadian Journal of Cardiology* 37(8): 1156–58. doi:10.1016/j.cjca.2021.03.003.

Yi, S.E., V. Harish, J.M. Gutierrez, M. Ravaut, K. Kornas, T. Watson et al. 2022. Predicting Hospitalisations Related to Ambulatory Care Sensitive Conditions with Machine Learning for Population Health Planning: Derivation and Validation Cohort Study. *BMJ Open* 12(4): e051403. doi: 10.1136/bmjopen-2021-051403.

About the authors

Laura C. Rosella, PhD, is the site director at ICES UofT and an associate professor and education lead at the Temerty Centre for Artificial Intelligence Research and Education in Medicine, University of Toronto in Toronto, ON. Laura may be contacted by e-mail at laura.rosella@utoronto.ca.

Vinyas Harish is an MD/PhD candidate at the Temerty Faculty of Medicine and the Dalla Lana School of Public Health, University of Toronto in Toronto, ON.