Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Practice Forum

Methodology minute: a machine learning primer for infection prevention and control

Timothy L. Wiemken PhD, MPH, FAPIC, FSHEA, CIC a,b,c,d,* , Ana Santos Rutschman SJD, LLM c,d,e

a Saint Louis University, School of Medicine, Department of Health and Clinical Outcomes Research, St. Louis, MO
b Saint Louis University, School of Medicine, Department of Medicine, Division of Infectious Diseases, Allergy, and Immunology, St. Louis, MO
c Saint Louis University, Advanced Health Data (AHeaD) Research Institute, Center for Systems Infection Prevention, St. Louis, MO
d Saint Louis University, Health Innovation and Legal Preparedness Partnership, St. Louis, MO
e Saint Louis University, School of Law, Center for Health Law Studies, St. Louis, MO

The use of machine-learning and predictive modeling in infection prevention and control activities is increasing dramatically. In order for infection preventionists to make informed decisions on the performance of any particular model as well as to determine if the output of the model will be useful for their program needs, a suitable understanding of the creation and evaluation of these models is necessary. The purpose of this primer is to introduce the infection preventionist to the most commonly used machine-learning method in infection prevention: supervised learning.

© 2020 Association for Professionals in Infection Control and Epidemiology, Inc. Published by Elsevier Inc. All rights reserved.

MACHINE LEARNING, AN OVERVIEW

Machine learning is an umbrella term encompassing many algorithms used for assisting human understanding of large amounts of data.1 Often, machine learning is conflated with the terms “predictive” or “prescriptive” modeling; though they are not directly interchangeable. Machine learning is also commonly confused with “artificial intelligence,” a phrase with little agreement on definition. Simply put, machine learning can be considered a computational method, while general artificial intelligence could be considered a physical manifestation utilizing machine learning to perform a task(s).

Several subclasses of machine learning have been developed and used in medicine,2 but the 3 major groups utilized include: unsupervised, supervised, and reinforcement learning. These groups include different algorithms, each with their unique pros and cons for specific tasks. Here, we focus on supervised machine learning, as it is the most commonly used approach for predictive modeling in infection prevention and control. Since this is only a brief primer, reviews can be found elsewhere in the literature for further study.1

SUPERVISED MACHINE LEARNING

When the phrase machine learning is used, frequently the speaker is talking about supervised machine learning, an approach most synonymous with predictive modeling. Supervised machine learning is a method where an algorithm learns from available data to create a model, which is then used to predict the same outcome for new data. The learning process is called “training the model,” where various features (AKA variables) are provided to an algorithm. The prefix “supervised” means that an outcome is known to the model during training. For example, one may be interested in being able to predict if someone has a catheter-associated urinary tract infection (CAUTI) before a diagnostic test. In this example, a binary variable, CAUTI, is the outcome of interest and would be obtained from an electronic health record retrospectively. A multitude of features that the modeler thinks might explain the presence or absence of CAUTI are added to the model as well. Depending on the algorithm used, different mathematical computations are done to allow the computer to learn and compare the complex patterns of the features in patients with and without CAUTI. Next, new data without the outcome known (eg, prospective patients with unknown CAUTI status) are passed to this model, which will output a prediction of the presence or absence of CAUTI. Supervised machine learning is used regularly in infection prevention for prediction of various health outcomes such as Clostridioides difficile infection1 and other healthcare-associated infections,4,5 and most recently for development of vaccine candidates for SARS-CoV-2.6

* Address correspondence to Timothy L. Wiemken, PhD, MPH, FAPIC, FSHEA, CIC, Saint Louis University, Department of Health and Clinical Outcomes Research, Division of Infectious Diseases, Allergy, and Immunology, 3545 Lafayette Ave #411, St. Louis, MO 63104.
E-mail address: timothy.wiemken@health.slu.edu (T.L. Wiemken).
Conflicts of interest: None of the authors have conflicts of interest to report.

https://doi.org/10.1016/j.ajic.2020.09.009
0196-6553/© 2020 Association for Professionals in Infection Control and Epidemiology, Inc. Published by Elsevier Inc. All rights reserved.
HOW GOOD ARE THESE MODELS? PERFORMANCE EXPLAINED

Model performance is focused on several statistics representing how good it is at predicting the outcome. Care must be taken when evaluating performance statistics as there are many and they are often utilized inappropriately, particularly model accuracy. A case-in-point is a product vendor who creates a model to predict sepsis. They report the model to be 90% accurate. Although great at face-value, it is more important to look at the sensitivity, specificity, positive, and negative predictive values, rather than a grand statistic such as accuracy. This is because accuracy is calculated as the total correct predictions divided by the total predictions, not separating false positives and negatives. Furthermore, if the model was trained inappropriately, this accuracy statistic may be incongruent with the performance of the model in real life. For example, if the model was trained on a dataset where 90% of the patients did not have sepsis and 10% did, the model could simply report that every patient did not have sepsis and it would still be 90% accurate. It would miss that 10% with sepsis, and the purpose of the prediction – to predict what is called the “minority class,” or the outcome group that is less frequent.⁷

WHAT VARIABLES OR FEATURES ARE IN THE MODEL?

A complex part of building a supervised machine-learning model is “feature engineering”; creating and modifying variables (features in machine-learning language) for inclusion. It is not as simple of an approach as traditional explanatory regression models where only “clinically meaningful” or “biologically plausible” variables are added to the model (eg, age, gender, race, etc.). Those variables can certainly be added to a machine-learning model, but the outcome should not be expected to be superior to any traditional regression model with the same variables.⁸ With machine learning, the goal is not to identify the impact of variables on the outcome with a ratio or odds ratio, but rather to have a model that can accurately predict an outcome. The features used in the model are largely irrelevant as long as they improve the performance of the model.

Natural Language Processing is another umbrella term which encompasses a great deal of methods for dealing with textual data, such as text mining (extracting specific words or phrases from a corpus of text), topic modeling (defining topics from a corpus), and defining the sentiment of text, among a wide variety of other methods often used for machine-learning models. Natural Language Processing has proven useful in various studies, including for identifying healthcare-associated infections from notes in the electronic health record.⁹¹⁰

A KEY ISSUE AND POLICY IMPLICATIONS

It is critical to understand that machine-learning models are based on data that are often likely to reflect contextual biases, due to having been produced in high-resource settings such as university-affiliated hospitals.¹¹ Outside these settings, data are unlikely to properly account for diversity in clinical features of the patients or may not have the same variables available (eg, various laboratory values) for modeling. Since machine-learning models strictly learn patterns present in the data supplied to the algorithm, any biased input data will result in biased outputs and predictions. As we begin to rely more on computational predictions in all areas of medicine and health, we must recognize and start addressing the data gaps. Failing to do so can lead to racism, sexism, ageism, or other forms of discrimination.

CONCLUSIONS

The future will bring more and more electronic decision support tools using machine learning to our workday. Remembering that any computer model is created by humans and may be error prone underscores the need to implement any decision support tool with caution, ensuring the clinical aspects of any predictions are not disregarded.

References

1. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health. 2020;41:21–36.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
3. Oh J, Makar M, Fusco C, et al. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. Infect Control Hosp Epidemiol. 2018;39:425–433.
4. Ehrentraut C, Kvist M, Sparrelid E, Dalani S. Detecting healthcare-associated infections in electronic health records: evaluation of machine learning and preprocessing techniques. SMMB. 2014;2014:3–10.
5. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalani S. Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. Health Informatics J. 2018;24:24–42.
6. Liu G, Carter B, Bricken T, et al. Computationally Optimized SARS-CoV-2 MHC Class I and II vaccine formulations predicted to target human haplotype distributions. Cell Syst. 2020;11:131–44.e6.
7. Ishwaran H, O’Brien R. Commentary: the problem of class imbalance in biomedical data. J Thorac Cardiovasc Surg. 2020;1–2.
8. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
9. Jacobson O, Dalani S. Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany; 2016.
10. Tanushi H, Kvist M, Sparrelid E. Detection of healthcare-associated urinary tract infection in Swedish electronic health records. Stud Health Technol Inform. 2014;207:330–339.
11. PW III. Medical AI and contextual bias. Harvard J Law Technol. 2019;33:65–116.