Digestive Diseases and Sciences (2022) 67:4604–4605
https://doi.org/10.1007/s10620-022-07511-x

EDITORIAL

Real-World Guidance from Artificial Intelligence? Predicting Outcomes of Inflammatory Bowel Disease Using Machine Learning

Danny Con[1] · **Abhinav Vasudevan**[1]

Accepted: 29 March 2022 / Published online: 3 May 2022

Management of chronic illnesses, including inflammatory bowel disease (IBD), can be challenging. There are many factors that contribute to the overall disease course and treatment response. In an ideal world, it would be possible to identify individuals with a high risk of disease complications with the assurance that the chosen treatment has the highest chance of success and is most likely to achieve meaningful long-term improvement.

Though traditional medical research with randomized controlled trials evaluate whether therapy provides efficacy under highly controlled conditions, their applicability to real-world individual patients is limited given the strict entry criteria and meticulous monitoring required in such studies. To address the increasing complexity of medical decision-making due to the explosion of available treatment options for chronic diseases, more sophisticated approaches are evolving that will facilitate more dynamic and potentially more individualized assessments of response and likely outcomes, termed "precision medicine," a concept that is gaining traction [1]. With the advent of electronic medical records, big data, and information technology rapidly dominating patient care, machine learning and artificial intelligence approaches are gaining increasing recognition for their potential ability to meet the needs of modern challenges in precision medicine [2, 3].

In this issue of *Digestive Diseases and Sciences* [4], Zand et al. analyzed a database of both privately and publicly insured individuals who are broadly representative of the US population in order to create a training and validation cohort for patients who held continuous insurance initially during 2015–6 and then in 2016–7 for the latter. Several machine learning algorithms including neural networks, random forests, and regularized regression were trained to predict the probability of adverse health outcomes, including IBD-related hospitalizations, IBD-related surgery, long-term corticosteroid use, and the initiation of biologics over the subsequent 12 months. A total of 108 predictive variables ("features") were used to construct the models. The training and validation cohorts each contained over 69,000 patients. In the validation cohort, 4.1% of patients required hospitalization, 2.9% required surgery, 13% needed long-term corticosteroids, and 17% required biologic therapy. The performance of all models was judged as at least "fair," with the most accurate models being the Random Forest and LASSO regression models, with area under the curve (AUC) values ranging from 0.7 to 0.92 for all outcome predictions. The relative importance of different predictive variables appeared to vary between the diverse models and outcomes, although IBD-related surgeries and long-term steroid use were strong predictors of IBD-related hospitalizations for LASSO and Ridge regression, whereas measures of health care utilization appeared to be strong predictors in the random forest model.

This study highlights key concepts and limitations in applied machine learning. Although machine learning models can accurately predict outcomes, the ideal algorithm best adapted to clinical medicine has not emerged given the respective benefits and trade-offs. For example, neural networks are highly accurate but prone to over-fitting, require large amounts of data to train, and utilize a "black-box" approach where the relationships between predictors and outcomes are not easily discernible for appraisal. In contrast, traditional regression approaches can produce simple bedside scores and are intuitive, yet perhaps lack the ability to model or predict complex and nonlinear data [4]. Further, it is unclear how the evaluated models compare with standard statistical models (for example, unregularized logistic regression). Moreover, cost–benefit analysis of implementing a potentially costly neural network algorithm against a cheaper regression model may be required. It is likely that

✉ Abhinav Vasudevan
abhinav.vasudevan@monash.edu

[1] Department of Gastroenterology and Hepatology, Eastern Health Clinical School, Box Hill Hospital, Monash University, 8 Arnold Street, Box Hill, VIC 3128, Australia

the optimal model will depend on the clinical question at hand while being subject to the trade-offs between simplicity, accuracy and consistency.

This study also highlights the importance of model validation and the risk of "over-fitting." The authors appropriately tested their candidate models from a different cohort of patients used in the derivation of the respective models. This is one form of model validation that seeks to test a model's ability to generalize to unseen data. The concept of over-fitting refers to the reduction in the performance of an algorithm when tested on a dataset external to the dataset used for model derivation that reflects the external validity or generalizability of the results. This performance reduction is due to the ability of sophisticated algorithms to learn relationships within a dataset that exist only due to chance or coincidence that would not be present in a different dataset. Although the authors do separate their database into a derivation and validation cohort to strengthen their findings, large databases can be inherently different from another depending on region and service provider, differences in coding practices, and variations in the manner in which data are collected. It is therefore expected that such predictive models will require validation studies among numerous external cohorts before their use in clinical practice can be advocated [5].

Model evaluation is another important consideration in applied machine learning research. Most studies assess model diagnostic performance using receiver operator characteristic (ROC) curves. Though the AUC provides a summary measure of a model's overall accuracy, this is unlikely to have the requisite discrimination needed to select the most appropriate model given clinical variables. Other factors, such as sensitivity, specificity and positive and negative predictive values, can be more informative [6]. For instance, screening tests for cancer need to be highly sensitive in order to avoid missing an undiagnosed malignancy. Although the authors provide sensitivity and specificity estimates, direct comparison of the models requires an understanding of the clinical question at hand; for example, when predicting IBD-related surgical risk, the health and economic cost of not operating on a patient who requires surgery (false negative) needs to be weighed against the cost of mistakenly performing surgery in a well individual (false positive). If the important clinical goal is to prevent escalation to surgical management, models should be optimized for their sensitivity (rather than overall AUC). Of note, none of the tested models had a sensitivity of $>75\%$ for predicting the need for surgery, reinforcing the principle that model evaluation should be carefully tailored to the clinical problem at hand so as to ensure that actual patient and societal outcomes are improved.

Despite the limitations of the current state of medical machine learning research, the use of machine learning algorithms and artificial intelligence is likely to continue to increase. These techniques can provide highly useful information used to assist clinicians in predicting outcomes using large datasets that are representative of adults with IBD as demonstrated by the study by Zand et al. [4]. These may eventually allow superior predictions of response to therapy over time and inform important management decisions that are tailored to the individual. Future research should adhere to best up-to-date practices in computer science research with appropriate model development and validation techniques. Clinicians should be mindful of the potential limitations of these methods when interpreting and applying the results to their clinical practice, especially as they may not be interpreted in the same manner as traditional statistical methods. Artificial intelligence and machine learning, some of the most exciting developments in the medical field, will grow in importance and influence on medical practice in the decades to come.

Author's contribution DC and AV were involved with the conception, design and writing of the manuscript.

Declarations

Conflict of interest Nil.

References

1. Denson LA, Curran M, McGovern DPB et al. Challenges in IBD research: precision medicine. *Inflamm Bowel Dis* 2019;25:S31-s39.
2. Chen H, Sung JJ. Potentials of AI in medical image analysis in gastroenterology and hepatology. *J Gastroenterol Hepatol* 2020;36:31–38.
3. Kohli A, Holzwanger EA, Levy AN. Emerging use of artificial intelligence in inflammatory bowel disease. *World J Gastroenterol* 2020;26:6923–6928.
4. Zand A, Stokes Z, Sharma A, van Deen WK, Hommes D. Artificial intelligence for inflammatory bowel diseases (IBD); accurately predicting adverse outcomes using machine learning. *Dig Dis Sci*. (Epub ahead of print). https://doi.org/10.1007/s10620-022-07506-8.
5. Le Berre C, Sandborn WJ, Aridhi S et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology* 2020;158:76-94.e72.
6. Moons KG, Altman DG, Reitsma JB et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73.