Machine Learning in Emergency Medicine: Keys to Future Success

A related article appears on page 184.

An era for artificial intelligence has arrived for emergency medicine. In the systematic review by Kareemi et al.[1] published in this issue of *Academic Emergency Medicine*, the authors evaluate the performance of machine learning (ML) models versus standard care (e.g., clinical decision rules, provider judgment) in emergency medicine across a variety of clinical scenarios and outcomes. The systematic review concludes that ML has superior performance in almost all tasks, but also calls attention to several widespread shortcomings including limited adherence to reporting guidelines and the lack of evaluation through interventional trials. These findings highlight the need for a new phase in clinical decision support (CDS) for emergency care with research and practice focused on integrated, ML-driven CDS systems that are usable, interpretable, and effective. In this commentary, we review key concept areas for enhancing the performance, promoting the adoption, and studying the impact of ML within emergency medicine (Figure 1). We also discuss the interpretation and application of ML studies and projects, dividing key concepts into two domains: intrinsic—elements of the model and its task-based performance—and extrinsic—the ability for the model to achieve a desired objective with respect to patient care.

WHAT IS ML?

Here, we first introduce readers to core principles that define ML and provide contrast to traditional statistical approaches. We focus specifically on "supervised" ML, where an algorithm is trained to recognize relationships between an outcome (e.g., mortality, length of stay) and any number of independent variables ("covariates" in statistical language and "features" in ML). This contrasts "unsupervised" learning, where data patterns are learned without attention to outcome prediction and is outside of the scope of this piece. One of the most common ways to investigate associations is via regression techniques, which include linear and logistic methods. Logistic regression has become ubiquitous because it efficiently weighs the relative contributions of multiple variables in an easily interpretable way.[2] Logistic regression is used in classification problems (i.e., predicting an outcome) and is a bridge between traditional statistics and ML approaches. Supervised ML algorithms all address the same fundamental prediction task, but use different underlying methods to derive their predictions. ML expands on traditional statistical approaches using general-purpose learning algorithms with minimal assumptions about the data-generating process to find patterns in rich and unwieldy data. ML models are therefore often employed when the focus is model accuracy; the data are multimodal or nontraditional (i.e., text, images, connectivity graphs); computational constraints exist; or when the goal is to capture complex, nonlinear relationships.

Intrinsic Performance

Critical assessment of ML algorithms requires understanding of model strengths and weaknesses, transparency in presentation, and methods to maintain deployed algorithms over time.

Feature Discovery

The success of ML algorithms generally depends on automated methods of data representation that can

The authors have no relevant financial information or potential conflicts of interest to disclose.
Supervising Editor: Jeffrey A. Kline, MD.

ACADEMIC EMERGENCY MEDICINE 2021;28:263–267.

doi: 10.1111/acem.14189
ISSN 1553-2712

uncover the different explanatory factors of variation behind the data. Conceptually, ML "features" and statistical "covariates" are synonymous when using traditional data types like continuous variables representing laboratory values. In these scenarios, however, ML models are readily able to expand on the logistic regression rule-of-thumb of limiting modeling to approximately one feature per 10 outcomes because they are much more resilient to overfitting.[2,3] Rapid growth in the number of variables highlights the need for parallel advances in interpretability and implementation, both of which will be discussed later.

Unlike traditional statistical regression models, ML models are able to look for nonlinear relationships in data. This advance is critical because not all interactions can be captured by repeatedly adding weighted variables together. For example, decision trees enable "OR" logic, like "does this patient have known CAD OR a smoking history." Exploration of nonlinear relationships is likely to highlight underappreciated clinical variables for further study.

Beyond using traditional covariates to directly make predictions, ML is also able to achieve one further level of abstraction called feature or "representation" learning, where the algorithm is asked to make predictions after first learning how to best look at the data. Representation learning has facilitated key advances in clinical computer vision (e.g., finding blood on head CTs) and in massive-scale medical note processing (e.g., predicting hospital readmission from discharge notes). ML models thus enable researchers to break out of traditional approaches using small number of human-selected features and instead cast a wider net that may enable enhanced predictive capabilities and the discovery of new relational and potentially causal factors for outcomes.

Model Assessment

Adopting state-of-the-art models and methods for their development will enhance performance and generalizability of ML solutions in emergency care. ML is now composed of a dizzying array of steps including preprocessing, exploratory data analyses, feature selection, training, model selection, validation, testing, and hyperparameter searches, which, as noted by Kareemi et al., complicates critical assessment. Expansion of and adherence to reporting guidelines and standards will play a key role for ML going forward. Recently, the EQUATOR network announced a planned expansion of the "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD) guideline for ML.[4] Studies reporting artificial intelligence interventions should adhere to CONSORT-AI or SPIRIT-AI guidelines.[5]

The MI-CLAIM checklist highlights metrics for evaluating algorithm performance addition to those for model clinical utility.[6] For the former task, area under receiver operating characteristic (AU-ROC, otherwise known as c-statistic), which is synonymous to the area under a model-derived sensitivity, 1-specificity curve is most commonly used, but it is important to recall caveats to its use. For example, in studies where one of the outcomes (e.g., mortality) is rare, it is more useful to present data from precision (i.e., positive predictive value)-recall curves.[7] Both these benchmarks capture model "discrimination," which says that patients at higher risk for an outcome should have a higher predicted risk. In contrast, model "calibration" provides information on the reliability of risk estimates for individuals, matters significantly for patient decision-making, and is often not reported.[8,9] Model clinical utility is also dependent on benchmarks that can be applied in the interpretation at the level of individual patients, namely, positive/negative predictive values, numbers needed to treat/harm, sensitivity, and specificity.[6]

Algorithmic Bias

Research has revealed troubling examples in which the reality of algorithmic decision making falls short of our expectations of impartiality and freedom from bias. Some algorithms have been shown to replicate and even amplify human biases, particularly those affecting protected groups.[10] In some cases, clinical implementation of race is explicit, such as in GFR estimation.[11] Elsewhere, bias can be an emergent property of missing types of data like inability to control for socioeconomic or health determinants. Finally, the use of proxy outcomes, like cost of care standing in place of health needs can lead to unintended and deeply problematic consequences.[12] Subgroup testing (e.g., gender and/or race) should be a core component of model assessment as should careful consideration of outcome choices.

Maintenance

As ML models are transitioned from individual benchmarking studies to longitudinally deployed tools, new standards for algorithmic stewardship will be needed.[13] There is mounting evidence that algorithms may experience "calibration drift" over time as features and outcomes change distribution over time.[14] Emergency medicine researchers, data scientists, and others interested in implementing ML models in the real world must begin to plan for the full life cycle of models, which will require funding, expertise, and periodic model reevaluation. Focus on model parsimony, that is, minimizing the number of features required to execute the prediction task, is likely to help this aim as a simpler model is easier to upkeep, but needs to be balanced against a desire for high performance.

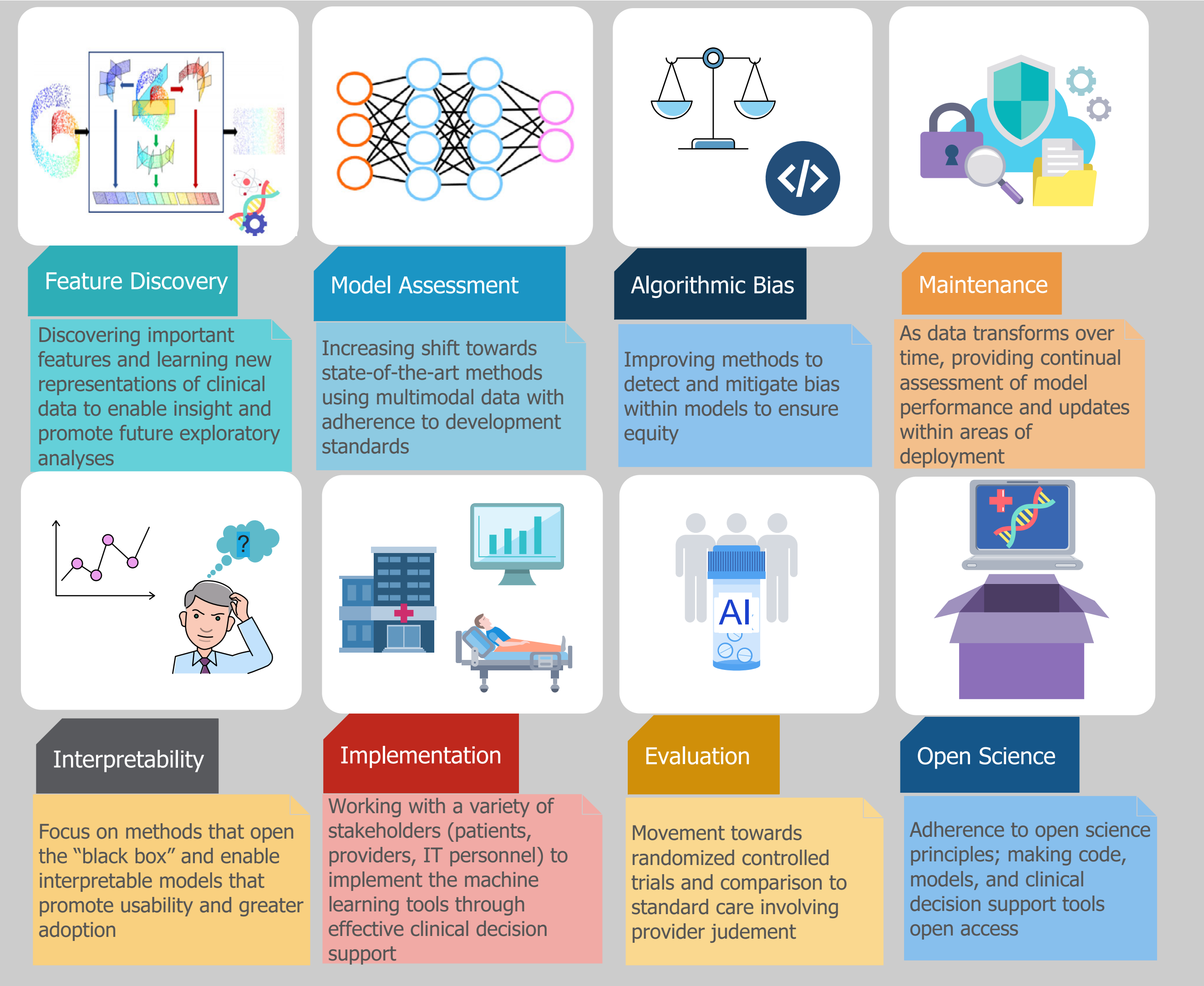
EXTRINSIC PERFORMANCE

The success of ML in emergency care requires implementation, measurement of impact on patient care, and dissemination to the larger community of healthcare providers.

Interpretability

Linear regression models have stood the test of time in part because they are easily interpreted and have known limitations. With the complexity of ML approaches comes complexity in interpretation—model predictions themselves are simple, but how they arrived at those findings are not. There is a fundamental difference in accepting the analysis of a field expert (e.g., a teleneuroradiologist interpreting an MRI) and doing the same for an algorithm, even though the requesting practitioner may find the study itself opaque.

Intepretability is the goal of opening the "black box" of artificial intelligence. Most ML tools have a mechanism to indicate importance of features in determining a model outcome, which provides some sense of algorithm underpinnings. Unfortunately, these rarely other the tangibility found in simple regressions.



Feature Discovery

Discovering important features and learning new representations of clinical data to enable insight and promote future exploratory analyses

Model Assessment

Increasing shift towards state-of-the-art methods using multimodal data with adherence to development standards

Algorithmic Bias

Improving methods to detect and mitigate bias within models to ensure equity

Maintenance

As data transforms over time, providing continual assessment of model performance and updates within areas of deployment

Interpretability

Focus on methods that open the "black box" and enable interpretable models that promote usability and greater adoption

Implementation

Working with a variety of stakeholders (patients, providers, IT personnel) to implement the machine learning tools through effective clinical decision support

Evaluation

Movement towards randomized controlled trials and comparison to standard care involving provider judement

Open Science

Adherence to open science principles; making code, models, and clinical decision support tools open access

Figure 1. Keys to the future success of machine learning in emergency medicine.

Significant leaps have been made in complex model interpretation, enabling both population-level and patient-level interpretation the key drivers of risk.[15,16] The MI-CLAIMS checklist includes requirements regarding model feasibility and interpretability.[6] Intrepretability also plays a key role in implementation.

Implementation

Dissemination and implementation (D&I) sciences are a growing force in emergency medicine just as they are beginning to contend with a new wave of ML tools.[17,18] D&I science serves as a reminder of the significant effort required beyond model development. ML will need to have a meaningful point of entry into the clinical workflow and accessible interpretation, while not demanding either too much attention or monopolizing computational resources. There is currently a significant barrier to entry for ML deployment as the financial incentives of health systems need to align with EHR vendors to purchase ML packages or build out their own data infrastructure. More fundamentally, ML will confront "the realities of entrenched institutions," defined in part by the decision-making authority of medical providers within the health care systems.[18] D&I sciences provide a framework for the engagement of provider, patient, organizational, and policy stakeholders.

Evaluation

Kareemi et al. find that ML outperforms usual care, as defined by clinician judgment, clinical decision tools, and triage-based scores. As with prior studies, they find that CDS tools are rarely compared to provider judgement.[19] We propose that new models be designed with eventual comparison to provider judgment in mind. Kareemi et al. also note a paucity of randomized trials rigorously assessing for meaningful differences in patient-centered outcomes. Multiple examples of interventional trials for CDS and AI, however, have emerged.[20,21] We anticipate rapid expansion of interventional trials with fewer barriers to implementation and the emergence of novel automated methods of EHR randomization.

Open Science

Open science, a movement that promotes sharing of both primary data and source code, is a promising initiative to enhance reproducibility. Part of the open science movement is the recognition that data management and analytic decisions have critical implications for interpretation and that computing workflows need to follow the same practices as lab projects and notebooks, with organized data, and documented steps. Adherence to these principles is critical for the success and transparency of ML in emergency medicine. Unfortunately, few studies published in widely accessed emergency medicine journals meet this bar.[22] Researchers, funders, and journals will need to align on standards and expectations for data and algorithm sharing.[6]

CONCLUSION

Kareemi et al. highlight the promise of ML. Realizing ML's potential in emergency care is a multifactorial challenge motivated by the potential for higher-quality, more efficient patient care. ML tools leverage many of the same core principles of traditional statistical approaches, while relaxing limitations on the number of variables under study, varieties of input data, and the types of relationships between variables. EM, a field defined by high volume, acuity, and flexibility, is poised to participate in this paradigm shift, informed both by the intrinsic (features, models, bias, maintenance) and by the extrinsic (interpretability, implementation, evaluation, open science) factors.

R. Andrew Taylor, MD, MHS
(richard.taylor@yale.edu)
Adrian D. Haimovich, MD, PhD
Yale Department of Emergency Medicine, Yale School of Medicine, New Haven, CT

References

1. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine learning versus usual care for diagnostic and prognostic prediction in the emergency department: a systematic review. Acad Emerg Med 2020;27.
2. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med 2011;18:1099–104.
3. Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30.
4. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577–9.
5. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ 2020;370:m3164.

6. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320–4.
7. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10: e0118432.
8. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA 2018;320:27–8.
9. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:1–7.
10. Mullainathan S, Obermeyer Z. Does machine learning automate moral hazard and error? Am Econ Rev 2017;107:476–80.
11. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. Mass Med Soc 2020;383:874–82.
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–53.
13. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. JAMA 2020;324:1397–8.
14. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Informat Assoc 2017;24:1052–61
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. Cambridge: The MIT Press, 2017:4765–74.
16. Haimovich AD, Ravindra NG, Stoytchev S, et al. Development and validation of the quick COVID-19 severity index: a prognostic tool for early clinical decompensation. Ann Emerg Med 2020;76:442–53.
17. Bernstein SL, Stoney CM, Rothman RE. Dissemination and implementation research in emergency medicine. Acad Emerg Med 2015;22:229–36.
18. Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Int Res 2019;21:e13659.
19. Schriger DL, Elder JW, Cooper RJ. Structured clinical decision aids are seldom compared with subjective physician judgment, and are seldom superior. Ann Emerg Med 2017;70:338–44.
20. Wilson FP, Shashaty M, Testani J, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. Lancet 2015;385:1966–74.
21. Yao X, McCoy RG, Friedman PA, et al. ECG AI-guided screening for low ejection fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. Am Heart J 2020;219:31–6.
22. Taylor RA, Haimovich AD, Horng S, et al. Open science in emergency medicine research. Ann Emerg Med 2020;76:247–8.