Digestive Diseases and Sciences (2019) 64:2061–2064
https://doi.org/10.1007/s10620-019-05698-0

EDITORIAL

Improving Acute GI Bleeding Management Through Artificial Intelligence: Unnatural Selection?

Neil Sengupta[1] · **David A. Leiman**[2]

Published online: 7 June 2019

Gastrointestinal bleeding (GIB) is the leading digestive disorder responsible for hospitalizations in the USA [1], a statistic not expected to change in the near future given the current shift of care toward more medically complex and anticoagulated patients, who are at increased risk for GIB. Accurate risk stratification of patients with GIB at initial presentation can facilitate improved triage efficiency and superior allocation of hospital-based resources. The ideal risk stratification tool should have both high positive and negative predictive values, which would result in low-risk patients' discharge for outpatient follow-up and early endoscopy in those with high-risk predictors. A health system embracing this model would likely reduce costs while potentially improving meaningful clinical outcomes such as overall mortality and hospital length of stay.

A variation on this theme is already being used in clinical practice, with multiple GIB risk stratification tools available, including the Glasgow Blatchford Score (GBS) [2]. The GBS was developed using a logistic regression model to predict the need for hospital-based intervention in upper GIB. Although accurate in identifying low-risk patients who can potentially be discharged early from the emergency department (ED) (GBS ≤ 1), the score is infrequently used in health systems to guide decision making [3]. Despite the evidence base and guideline support for its use, several challenges to widespread implementation exist. First, there are inherent limitations of the model such as relatively low specificity [4], with consequent missed opportunities to identify many low-risk patients with upper GIB, and inaccurate identification of patients at high risk for mortality and need for urgent endoscopic intervention. Moreover, risk score determination typically requires the use of a separate calculator that is often external to the electronic health record (EHR) environment. Finally, while gastroenterologists may be those most familiar with scoring tools, initial triage decisions do not always involve their input. Given the manifest burden of GIB, a new approach is needed.

Machine learning (ML) refers to how computers can integrate and analyze large amounts of data in order to identify useful patterns and associations. With their prodigious processing capacity and power, this technique can assist in detecting novel relationships between inputs or variables in order to improve the prediction of clinical outcomes that may not be predicted by conventional modeling such as linear or logistic regression [5]. Supervised ML refers to the identification of a specific outcome based on a variety of individual features. A computer model is trained to map individual features to an outcome and subsequently tested on a dataset that was not used to develop the model. Artificial neural networks (ANN) are one type of ML model that can be used to evaluate complex, nonlinear relationships between individual features. Inputs are transformed into several layers of weighted hidden nodes, leading to an output. ML algorithms have already been used to predict important clinical outcomes in hospitalized patients such as occurrence of sepsis [6], hospital readmission [7], and *C. difficile* infection [8] using structured data (inputs and outcomes) residing in the EHR. In these studies, the ML-based predictions were more accurate than corresponding logistic regression models [9]. Although promising, data regarding prospective validation and real-world utilization of these models are limited.

GIB may serve as an ideal case for ML in gastroenterology (Fig. 1). In addition to a need for better risk stratification, there are numerous defined predictor variables for GIB

✉ Neil Sengupta
nsengupta@medicine.bsd.uchicago.edu

David A. Leiman
david.leiman@duke.edu

[1] Section of Gastroenterology, Hepatology, and Nutrition, University of Chicago Medical Center, 5841 S Maryland Avenue, MC 4076, Chicago, IL 60637, USA

[2] Division of Gastroenterology, Duke University School of Medicine, 200 Morris Street, Room #6524, Durham, NC 27701, USA

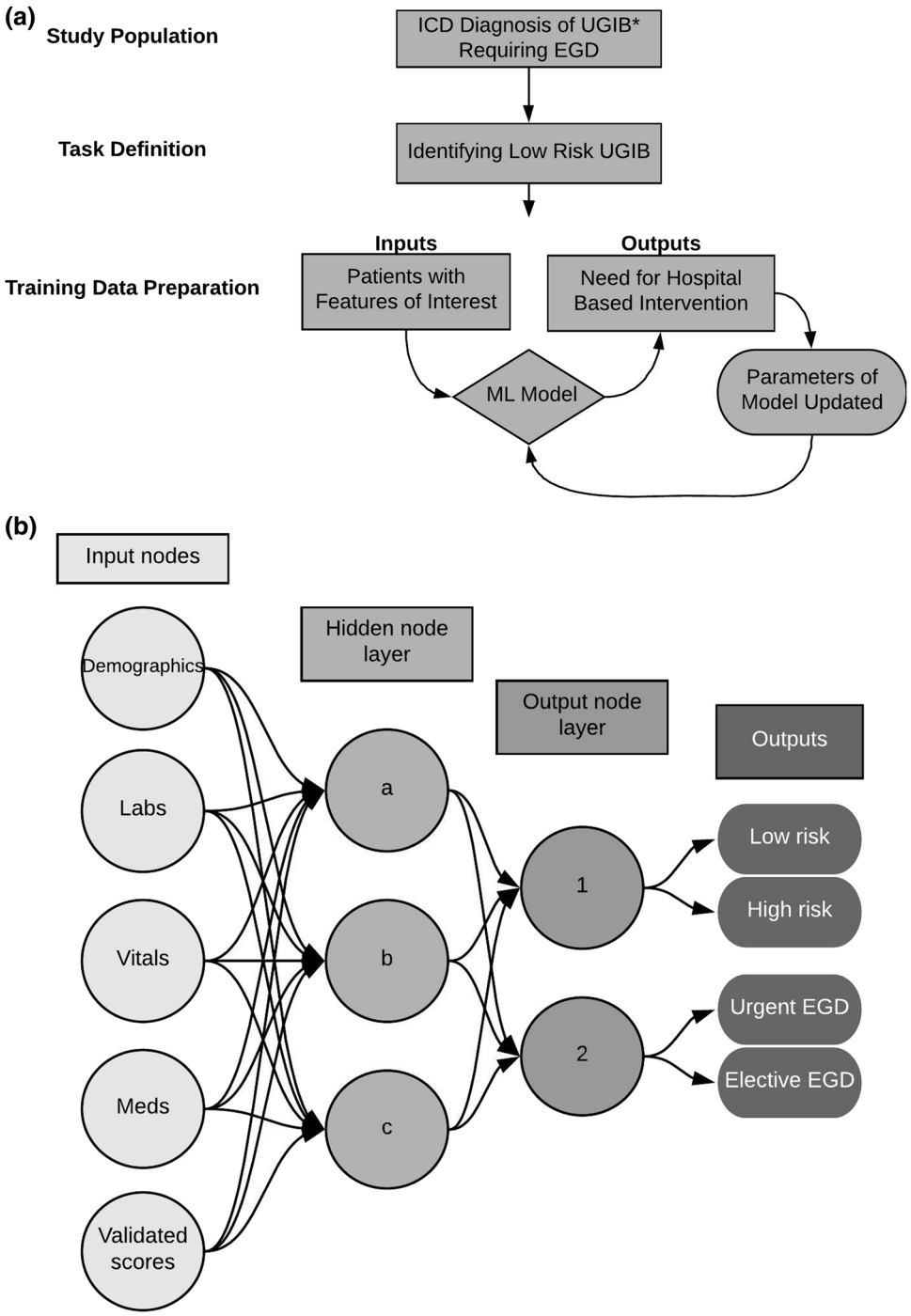

Fig. 1 Building a machine learning model for GIB. **a** An example task is to predict patients with upper gastrointestinal bleeding (UGIB) who are at low risk for a hospital-based intervention. We identify the inclusion population, which consists of patients presenting with UGIB (identified via billing codes) who underwent upper endoscopy (EGD). Data preparation relies on inputs (individuals with distinguishing features) and outputs (whether the patient required a hospital-based intervention). Training data are run through a machine learning (ML) model and evaluated for prediction. The predicted outcome is compared to the ground-truth label (whether patients experienced a hospital-based intervention). If the prediction is incorrect, the model parameters are updated to make the model more likely to make a correct prediction. After training is complete, the model is tested with unlabeled data with predictions compared to ground truth. **b** In this case, the ML model is an artificial neural network (ANN). Input nodes consist of demographics, laboratory values, initial vital signs, medications, and established prediction scores such as the GBS. Input data are transformed into weighted, hidden node layers, leading to the prediction of an output, *i.e.*, whether a patient is at low risk for hospital-based intervention. Data from ANNs can also be used to predict a quantitative response. A theoretical example would be percent reduction in rebleeding risk with early endoscopic intervention. *GIB* gastrointestinal bleeding, *ICD* International Classification of Diseases, *UGIB* upper gastrointestinal bleeding, *EGD* esophagogastroduodenoscopy, *ML* machine learning, *ANN* artificial neural network, *GBS* Glasgow Blatchford Score

embedded within the EHR that are not optimally leveraged with the existing risk scores. These include structured data such as demographics, comorbidities, admission medications/doses (e.g., anticoagulant drugs), initial laboratory findings, vital signs, as well as unstructured data such as initial symptoms. Furthermore, well-defined, high-risk outcomes needed to train an ML model such as need for blood transfusion, endoscopic intervention, and mortality exist in the EMR. To realize the stated goals of ML-based prediction—namely improved identification and prognostication as well as more efficient resource allocation leading to improved outcomes—more clinical data and accumulated ML experience are needed.

As a result, Shung et al. [10] in this issue of *Digestive Diseases and Sciences* provide a useful and timely summary of the available evidence of ML techniques used to predict the outcomes of GIB. The results of this systematic review highlight the promise of this technology, but more importantly identify areas for future work. The authors only included studies evaluating the utility of an ML-based tool in predicting outcomes of mortality, rebleeding, or hemostatic intervention with respect to GIB. Appropriately, they excluded studies that did not include an internal or external validation dataset. A total of 14 observational studies that used a total of 30 ML models were included and discussed. Of note, only 11 studies compared ML to traditional non-ML tools to predict GIB, and no studies were deemed to have a low risk of bias. The median area under the receiver operating characteristics curve (AUROC) in validation (internal and external) datasets for a predefined outcome of mortality, hemostatic intervention, or rebleeding was 0.87 (range 0.40–0.98) with the ML tools. In studies of upper GIB, ML tools were superior to the GBS and other clinical risk scores for predicting mortality.

The authors should be commended for conducting this study and for synthesizing these data. Based on the data presented, ML models (particularly ANNs) showed promise in accurately predicting high-risk outcomes of GIB compared to traditional risk scores. However, a significant limitation of the included studies is the small number of subjects per study (147–2380), numbers that are orders of magnitudes less than what is usually required to accurately train computer vision ML models. Therefore, the high AUC values of the cited ML models can be from over-fitting of a training dataset, which would complicate replication of the results. Shung et al. show that prior to clinical implementation of an ML model, much larger training and testing databases, in addition to prospective validation of these tools in external populations, are needed. The disconnect between the enthusiasm for ML and the current state of artificial intelligence in clinical practice is evident.

A variety of challenges exist to routinely implement ML models in clinical care settings. Unless there are data comparing not just performance characteristics but actual meaningful endpoints, such modeling will remain an academic exercise. Among other important barriers is the obvious vulnerability of ML models to misleading inputs; there will need to be means to check the accuracy and veracity of clinical data, since this will form the foundation upon which ML models are generated and updated. This in turn will require curated and structured inputs relevant to GIB, and agreement on standard definitions for low- and high-risk outcomes of upper and lower GIB. One way to achieve standardization and realize the goal of high-quality care may be through the development of GIB-specific quality metrics that should be derived from guidelines and could help ensure that the data used in ML algorithms are reliable. As a result, similar structured data could be shared between institutions, practices, and health systems. The resulting records would facilitate the large training and testing datasets required to build accurate ML tools.

Nonetheless, even if the data used in ML algorithm development are trustworthy, there may be concern about the "black box" outputs generated from complex ML algorithms such as ANNs. Moreover, using the same predictors included in traditional scores such as the GBS and simply incorporating them into a complex ML algorithm is unlikely to add much to the existing standard risk stratification tools. Rather, ML provides an opportunity to study innovative features related to GIB prediction such as the initial trajectory of vital signs, laboratory parameters, and/or interaction between multiple medications. Finally, since the optimal approach to incorporating the data received from ML models into practice remains unknown, this should be an area of particular focus and study going forward. As described by Shung et al., an essential next step for this technology will be ensuring a close collaboration between those programming algorithms and those using them regularly, in order to ensure maximal impact. It is likely that for optimal benefit, ML models should be directly embedded into the EHR in order to alert and update physicians in real time about individual patients' risk level, and serve as an adjunct to clinical decision making.

The synthesis provided by Shung et al. is valuable to clinicians aiming to understand the current state of machine learning, as well as its potential for the effective and efficient management of acute GIB. While still maturing, this technology has evident promise. Future work in the development of computational models will rely on creating standardized and trustworthy data elements and prospective testing in large trials. Any eventual incorporation into clinical practice likely depends on a greater understanding and acceptance of ML assistance in clinical care. But, if developed collaboratively between groups and health systems with meaningful clinical outcomes in mind, ML has the potential for positively impacting patient management and care delivery.

Compliance with Ethical Standards

Conflict of interest The authors declare they have no conflict of interest.

References

1. Peery AF, Crockett SD, Murphy CC, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the united states: update 2018. *Gastroenterology*. 2019;156:254–272 e11.
2. Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for upper-gastrointestinal haemorrhage. *Lancet*. 2000;356:1318–1321.
3. Leiman DA, Mills AM, Shofer FS, et al. Glasgow blatchford score of limited benefit for low-risk urban patients: a mixed methods study. *Endosc Int Open*. 2017;5:E950–E958.
4. Stanley AJ, Laine L. Management of acute upper gastrointestinal bleeding. *BMJ*. 2019;364:l536.
5. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930.
6. Henry KE, Hager DN, Pronovost PJ, et al. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015;7:299ra122.
7. Shameer K, Johnson KW, Yahi A, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort. *Pac Symp Biocomput*. 2017;22:276–287.
8. 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.
9. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
10. Shung D, Simonov M, Gentry M, et al. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: a systematic review. *Dig Dis Sci.* (Epub ahead of print). https://doi.org/10.1007/s10620-019-05645-z.