EDITORIAL

‘No growth to date’? Predicting positive blood cultures in critical illness

Vincent X. Liu^{1*} and Jenna Wiens²

© 2020 Springer-Verlag GmbH Germany, part of Springer Nature

The rapid growth of healthcare predictive models has been fueled by the widespread availability of three essential elements: (1) large-scale digital and electronic health record (EHR) data; (2) off-the-shelf, robust machine learning (ML) algorithms; and (3) powerful, low-cost computing platforms [1]. Together, these changes have fostered an ‘arms race’ towards superior predictive performance highlighting the latest algorithm, pipeline, or process. Concurrently, the complexity of model development has quickly outpaced the reporting capabilities of traditional scientific journal formats [2]. Thus, today, even experienced readers are stymied by the increasingly difficult task of assessing the value of a rapidly growing number of predictive models [3].

These challenges are likely to spur new approaches to evaluating and comparing healthcare predictive models. In the meantime, we continue to focus on assessing the value of a model within three key domains: (1) clinical utility, (2) validity, and (3) feasibility and sustainability (Table 1). While the listed items are not comprehensive, they can help identify key gaps and potential pitfalls.

To see this framework in action, we apply it to the study by Roimi et al. [4] which sought to predict bloodstream infections among patients with suspected infection in two independent ICU’s (Beth Israel Deaconess Medical Center in the US and Rambam Health Care Campus in Israel; BIDMC and RHCC). Using time-stamped EHR data consisting of medications, laboratory values, vital signs, treatments, and comorbidities, the authors implemented tree-based ML methods to identify patients

whose blood cultures, drawn after > 48 h in the ICU, would be positive with a clinically significant pathogen.

Two relatively novel features of this study were that rather than focusing on building a single model suitable for both sites, they created site-specific models that instead shared similar feature engineering (i.e., data pre-processing and variable identification) and modeling pipelines. Second, they designed a process that incorporated successive transformations, summary statistics, and/or time-series trends, such that each dataset expanded to include >7000 predictors. To prevent over-fitting, their final modeling approach included only the top 50 most predictive variables.

Among 31.7% of ICU patients with >48 h blood cultures at BIDMC, 6.4% had a positive culture. BIDMC-specific model discriminative performance (area under the receiver operating characteristic curves, AUROC) was excellent (test set: 0.89–0.90). Similarly, among 56.3% of patients with RHCC blood cultures, 15.9% had a positive culture with RHCC-specific test AUROC of 0.92–0.93. At a sensitivity threshold of 50%, specificity was $\geq 93\%$ with a low number needed to screen to detect one true-positive case of 1.2–3.

To evaluate the value of this predictive model, we begin by assessing clinical utility. Patients with bloodstream infection have high mortality, exceeding 25%, making them a prime target for care improvement [5, 6]. A substantial proportion (33%–42%) of patients undergoing blood culture is not on antibiotics [7]. In this study, in the 6 h following cultures, 21%–28% had not yet received antibiotics. Although not an identical population, sepsis patients benefit from earlier antibiotics, suggesting an opportunity to accelerate the initiation of appropriate antibiotic use among patients with a high predicted risk of blood culture positivity [8].

At the same time, unnecessary antibiotics are increasingly being recognized as a threat both at the patient and

*Correspondence: Vincent.x.liu@kp.org

¹ Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612, USA

Full author information is available at the end of the article

Table 1 Key domains and considerations for evaluating the value of a predictive model

Domain	Key summary questions	Characteristics that improve a prediction model's likelihood of success	Potential concerns if the domain is not well-optimized
Clinical Utility	Does the end-user understand the specific action that will be taken following the prediction? Are those actions targeting a clearly defined moment of care, patient population, and/or modifiable outcome?	Target clinical cohort undergoing prediction is clearly delineated Prediction triggers are well understood and feasible Clinical outcome is clinically important and modifiable Model end-user is defined and model use is within scope of practice Predictions and actions are comprehensible by routine end-users and patients Clear action that follows prediction is specified and realistic Pathway to default integration in healthcare workflow is available	Models may fail to improve, or even worsen, outcomes Predictions may be misapplied to patient groups or clinical decisions beyond what the model supports Clinical workload might increase resulting in inefficiency and clinician fatigue Translation timeline from model development to clinical use can be significantly delayed Clinician or patient uncertainty about model implications may stymie ongoing use of tool
Validity	Are key elements of model development—for example, the prediction interval, the feature engineering pipeline, and the training/validation experimental setup—appropriate for the prediction task and modeling approach? Are model performance and intelligibility adequately described, compared, and validated?	Prediction outcome, interval, and horizon are clearly stated Dataset sampling rates are clear Feature extraction and engineering are well-described and justified Modeling approaches are appropriate Test set is held-out and representative of intended use case Model performance is judged broadly, rather than solely focused on discrimination Model intelligibility or validation is adequate to assess performance Prospective or external validation is assessed and successful Potential for model to engender unintended consequences is considered and minimized	Models will fail to show comparable performance when applied to new settings, including external test sets or real-world use Although predictive performance appears better than comparators, incremental statistical improvement fails to generate meaningful clinical improvement Calibration is poor, particularly at the specific thresholds where clinical decisions are being made Low positive predictive values result in large proportion of false positive alerts
Feasibility and sustainability	How difficult will it be to instantiate the predictive model and ensure that its technical operation is robust and maintainable? How might the model positively and negatively impact clinical and technical workflows when deployed prospectively?"	Platform customizations for a single model are minimized Data feeding the prediction model are robust and have reasonable temporal stability Irregularities in data that drive prediction can be monitored over time Technical capability to update model into the future is available Outcomes evaluation and balancing measures are assessed prospectively Workload burdens produced by prediction are minimized and sustainable Unintended consequences of model in action are considered and mitigated	Technical deployment is highly customized and require skills beyond the capabilities of intended setting Models are overfit to workflow and documentation patterns that can radically change over time Risk alert-driven interventions negatively impact future model performance, without consideration for model re-fitting or re-calibration Technical and clinical burden are unsustainable, particularly if numerous models require highly customized platforms Only a single clinical outcome is tracked, without consideration for evaluating balancing measures that could point to unintended harm or disparities

population levels. A predictive model that lacks a strong negative likelihood ratio could inappropriately increase antibiotic exposure in certain patients depending on if and how negative scores are presented to clinicians. At a sensitivity threshold of 50%, the expected impact of this model's negative likelihood ratio (0.51–0.54) would only be small to moderate. Competing diagnostic tests such as procalcitonin and rapid pathogen assays are also entering mainstream clinical use and may offer alternate approaches to risk stratification [9, 10].

Together, these factors suggest this model has at least moderate potential clinical utility for patients not on

antibiotics at blood culture sampling. However, the precise question of how the end-user will specifically act on the predicted risk remains incompletely defined. Moreover, since the action will likely differ based on whether or not patients are already on antibiotics, evaluation within each of these subpopulations is important.

Assessing model validity is a significantly more challenging task. Despite the authors including a moderately high level of detail in the manuscript, key questions remain about the feature engineering and modeling pipeline. Why is the single strongest predictor of blood culture positivity—the duration between blood culture

sampling and last defecation—as much as fivefold more important than the next few variables? Indeed, defecation variables make up seven of the top 100 variables across sites. Similarly, why does the model incorporate variables not traditionally seen as clinically relevant (mean corpuscular hemoglobin concentration or the partial pressure of alveolar oxygenation divided by positive end-expiratory pressure)? Are these novel insights, proxy measures of other key predictors, or spurious findings?

In the modeling approach, why did the authors implement an unusual soft voting method using multiple tree-based methods, when routine approaches would favor increasing the number of trees within a single method? Why do across-site and simple logistic regression model performances degrade so significantly? Would a regularized linear model using the full feature set improve performance? Variability in data sampling might account for these differences, but could also make the models brittle in real-world use. In this case, incomplete insight into the approach could impair ‘live’ model validity in real-world settings.

Finally, we assess feasibility and sustainability. An advantage of site-specific model development is that the resulting predictions maximally leverage local patterns and data [11]. In theory, this makes them better suited to implementation within current clinical workflows. On the other hand, such predictive models may ‘overfit’ specific clinical workflows and EHR documentation processes making them difficult to compare between sites. In addition, if practices change, model performance could degrade significantly. Institution-specific models also require staff who can implement and maintain local algorithms, as well as data with enough samples and granularity to follow a similar modeling strategy. Further, as modeling approaches become more complex, addressing declines in future model performance and mitigating unintended consequences can become increasingly challenging.

Although this study highlights intriguing new ‘growth to date’ by leveraging routine clinical data to predict blood culture positivity in the ICU, key questions about the model’s clinical utility, validity, and feasibility/sustainability remain. These considerations should be used to inform prospective clinical testing, which remains the only way to evaluate true model value. The study also highlights the tremendous new potential that predictive models hold for improving critical care. Predictive tools are becoming commoditized; however, the challenges in assessing model value will remain. Considering

this changing landscape, new venues and modalities for assessing and comparing model development, performance, and value will become increasingly important for a successful future of predictive analytics in healthcare.

Author details

¹ Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612, USA. ² Division of Computer Science and Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, USA.

Funding

VXL is funded by The Permanente Medical Group and NIH R35GM128672.

Compliance with ethical standards

Conflicts of interest

The authors report no conflicts of interest.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 6 December 2019 Accepted: 24 December 2019

Published online: 21 January 2020

References

- Obermeyer Z, Emanuel EJ (2016) Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 375(13):1216–1219
- Beam AL, Kohane IS (2018) Big data and machine learning in health care. *JAMA* 319(13):1317–1318
- Chen JH, Asch SM (2017) Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med* 376(26):2507–2509
- Roimi M, Neuberger A, Shrot A et al. (2020) Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. *Intensive Care Med*. <https://doi.org/10.1007/s00134-019-05876-8>
- Valles J, Ferrer R (2009) Bloodstream infection in the ICU. *Infect Dis Clin North Am* 23(3):557–569
- Wisplinghoff H, Bischoff T, Tallent SM, Seifert H, Wenzel RP, Edmond MB (2004) Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clin Infect Dis* 39(3):309–317
- Tabah A, Koulenti D, Laupland K et al (2012) Characteristics and determinants of outcome of hospital—acquired bloodstream infections in intensive care units: the EURO-BACT international cohort study. *Intensive Care Med* 38(12):1930–1945
- Liu VX, Fielding-Singh V, Greene JD et al (2017) The timing of early antibiotics and hospital mortality in sepsis. *Am J Respir Crit Care Med* 196(7):856–863. <https://doi.org/10.1164/rccm.201609-1848OC>
- Ratzinger F, Haslacher H, Perkmann T et al (2018) Machine learning for fast identification of bacteraemia in SIRS patients treated on standard care wards: a cohort study. *Sci Rep* 8(1):12233
- Sinha M, Jupe J, Mack H, Coleman TP, Lawrence SM, Fraley SI (2018) Emerging technologies for molecular diagnosis of sepsis. *Clin Microbiol Rev*. <https://doi.org/10.1128/CMR.00089-17>
- Oh J, Makar M, Fusco C et al (2018) A generalizable, data—driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *Infect Control Hosp Epidemiol* 39(4):425–433