Title

Making the Improbable Possible: Generalizing Models Designed for a Syndrome-Based, Heterogeneous Patient Landscape

Journal

Critical Care Clinics, 39(4)

ISSN

0749-0704

Authors

Le, Joshua Pei

Shashikumar, Supreeth Prajwal

Malhotra, Atul

et al.

Publication Date

2023-10-01

DOI

10.1016/j.ccc.2023.02.003

Peer reviewed

HHS Public Access

Author manuscript

Crit Care Clin. Author manuscript; available in PMC 2024 April 01.

Published in final edited form as:

Crit Care Clin. 2023 October ; 39(4): 751–768. doi:10.1016/j.ccc.2023.02.003.

Making the Improbable Possible: Generalizing Models Designed for a Syndrome-Based, Heterogeneous Patient Landscape

Joshua Pei Le, BS[a], **Supreeth Prajwal Shashikumar, PhD**[b,1], **Atul Malhotra, MD**[c,1], **Shamim Nemati, PhD**[b,1,2], **Gabriel Wardi, MD, MPH**[c,d,*,2]

[a]School of Medicine, University of Limerick, Castletroy, Co, Limerick V94 T9PX, Ireland

[b]Division of Biomedical Informatics, University of California San Diego, San Diego, CA, USA

[c]Division of Pulmonary, Critical Care and Sleep Medicine, University of California San Diego, San Diego, CA, USA

[d]Department of Emergency Medicine, University of California San Diego, 200 W Arbor Drive, San Diego, CA 92103, USA

Keywords

Data science; Data missingness; Machine learning; Syndrome; Sepsis; Critical care

INTRODUCTION

Although machine learning (ML)-based clinical decision support has seen some successful implementations in radiology[1,2] and ophthalmology,[3,4] its overall presence in health care is modest when compared with its potential. This underutilization remains particularly true in the intensive care unit (ICU). A major hurdle to the widespread deployment of ML models has been their inconsistent performance as a result of several factors, including hospital-dependent operating procedures, patient demographics, and missing data.[5–7] These factors all contribute to data heterogeneity, so an ML prediction model (herein termed "ML model" or "model") trained on one dataset may exhibit degradation in performance when deployed on another,[8] resulting in inadequate *generalizability*. However, as noted by Futoma and colleagues, the colloquial use of the term generalizability in clinical ML literature is broad and not well-defined.[9] For clinical applications of ML models, a published hierarchy[10] describes it as a set of rules that may apply to internal, temporal, and external applications relative to the original training dataset. An internal application refers to using an ML model to the same patient cohort on which it was trained (eg, the same dataset at the same hospital). A temporal application denotes using this model at the same location but across a different time period. An external application refers to using this model at a separate location during any time period. The ideal model will be able to demonstrate similar levels of performance under any application, notably external ones.[11–14] This makes sense—it is

*Corresponding author. gwardi@health.ucsd.edu.
[1]Present address: 9500 Gilman Dr, La Jolla, CA 92093, USA.
[2]Authors equally contributed.

essential to verify that a model for clinical use can provide similar results for any group of patients, especially those it was not explicitly trained on.[9] Yet due to the current nature of ML and statistics, external generalizability approaches a limit as the applicable population grows because of increasing variation in workflow practices, point-of-care measurement devices, and population characteristics. As such, the current literature suggests that it is infeasible to develop a single universal prediction model. Therefore, our focus should shift toward more clearly defining the "conditions for use" of a model while improving its external generalizability *within* the intended use population. These "conditions for use" are ideally broad enough so that a prediction model can be impactful for as many patients as possible while minding the aforementioned limitations.

Patient-focused clinical predictive modeling can be classified as diagnostic versus prognostic.[15] Diagnostic tasks rely on a patient's "true state," which are typically defined by proxy criteria and construct via laboratory values, imaging, and physical symptoms, among others, to predict a clinical development (eg, sepsis). Prognostic tasks rely on the ordered tests and their results to quantify the probability of certain patient-centered outcomes (such as in-hospital mortality).[16] Both tasks have clinical utility, albeit at varying levels. Early diagnostic tasks may be able to assist clinicians administer timely interventions for a developing condition (eg, early and appropriate antibiotics for sepsis). Within the same context, prognostic tasks may help administrators optimize resource allocation for patients who are most likely to decompensate or are at risk for other adverse outcomes, including mortality.[15,17–19] In either case, existing studies investigate such tasks in critical care environments because the nature of its routine patient monitoring makes for data-rich electronic health records (EHRs) from which models can learn and make predictions.

In this review article, the authors investigate the current challenges of integrating ML models into critical care, using studies centered on early sepsis prediction as examples. The authors (1) explore clinical challenges with syndrome-based conditions, which are commonly diagnosed in the ICU; (2) clarify data science terminology surrounding these studies; (3) examine major barriers to generalizability; and (4) illustrate how current-day ML models address such obstacles via different methods of learning. The authors conclude with a discussion on areas for future research.

CHALLENGES WITH SYNDROMES

A syndrome can be defined as a recognizable complex of findings and symptoms that indicate a specific condition with a poorly understood cause.[20] A disease refers to a condition in which a causative agent or process results in a readily identifiable clinical and biological manifestation. Yet, with increased research and study, a condition that was formerly best described as a syndrome can be referred to as a disease. Indeed, Kawasaki *disease* was initially known as mucocutaneous lymph node *syndrome* because the underlying pathophysiology was uncertain and clinical manifestations were varied.[21] An increased understanding of the disease process later described clearly identifiable diagnostic features and treatment responses.

Although this distinction between a disease and syndrome is easily understood by physicians, syndromic conditions can prove challenging for ML models to identify and

predict. This is particularly true in the ICU because a multitude of prevalent syndromes (ie, sepsis and the acute respiratory distress syndrome) share similar physiologic and biological derangements. For example, a patient with decompensated heart failure can show vital signs and laboratory findings that mimic septic shock. In addition, critically ill patients are frequently comorbid for these syndromes. Clinicians use contextual clues, historical components, and physical examination findings to help differentiate between possibilities, but contemporary ML models struggle because historical context and examination findings are oftentimes not readily available[14] for such rapidly evolving patients. Complicating this further are multiple definitions for any given syndrome in sepsis, the condition can either be defined by Sepsis-3,[22] Severe Sepsis and Septic Shock Management Bundle (SEP-1),[23] or the Center for Disease Control and Prevention (CDC)[24,25] criteria.

CLARIFYING TERMS

Before we begin our review on generalizability and related ML studies, it is necessary to clarify terms and concepts that are commonly used in data science and relevant to its application in critically ill patients. Three recently published ML models, Artificial Intelligence Sepsis Expert (AISE),[11,12] Weight Uncertainty Propagation and Episodic Representation Replay (WUPERR),[13] and COnformal Multidimensional Prediction Of SEpsis Risk (COMPOSER),[14] and their respective studies are used as several examples throughout this article. Table 1 summarizes definitions and examples of relevant vernacular.

Missingness describes the presence of missing data and how it is handled, reported. This can negatively impact the predictive accuracy of a model and decrease its clinical utility.[8] However, not all missing data are created equally. There exist three primary classifications of missing data[26]: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). MCAR refers to randomly missing data that does not have any distinguishable pattern to it.[27–29] MAR refers to randomly missing data that may be associated with an underlying pattern.[27,30,31] For instance, in a hypothetical case where missing Glasgow Coma Scores (GCS) among trauma patients were more likely to be observed for older patients,[32] the mechanism is MAR. MNAR refers to the likelihood of a missing value to seem as a function of the value itself[27]—following the same hypothetical, the mechanism is MNAR if a missing GCS is known to be associated with mild brain trauma.[32] For prediction models, the MNAR mechanism is typically assumed because they use longitudinal data collection, where patterns of missingness are more likely to present (eg, healthy patients routinely have fewer tests as they are not indicated). The differences in how data go missing are key to understanding their handling,[27,31] which are more deeply explored in the section Approaches to Data Missingness.

At its core, ML consists of a set of parameters that are initialized with random weights. Training then takes place, where the model is exposed to a development cohort (eg, retrospective clinical data at Hospital A). This induces key changes in the weights of different rules. The idea is to give more bearing on the final output to rules that are mathematically determined to be “more important.” Prediction error (or the difference between the predicted outcome and the true patient outcome) will tell the model when it produces a correct or incorrect output, prompting changes to these rule weights, a process

known as supervised learning.[33] As it is exposed to a validation cohort (eg, retrospective clinical data at Hospital B), performance usually degrades because differences in the underlying data often necessitates different weights being given to different rules; this is the basis of the barriers to generalizability.[5]

ONE SIZE DOES NOT FIT ALL: THE GENERALIZABILITY PROBLEM

Generalizability was previously described as the ability for an ML model to perform similarly well on both development and validation cohorts within a well-defined intended use population. It would ideally maintain this high level of performance as it is applied to additional institutions that fit its "conditions for use." However, in order to do so, we must acknowledge the challenges it faces at different locales. These include explicit differences between institutions and missing clinical data. The following sections detail these problems in addition to potential methods that ML models might use to overcome them.

Heterogeneity in Health Care—Despite all its regulations, health care remains an area of great heterogeneity at various levels. Shashikumar and colleagues describe these levels in a recent *Behind the Paper*[5] on Nature Portfolio Health Community, starting with EHRs: different EHR vendors currently encode information in non-standardized formats which reduce their interoperability. The data are recorded using clinical instruments from different vendors which often use proprietary data processing methods with varying necessity for clinician verification. Local guidelines vary in their frequency for specific clinical measurements to be taken, but this should not be confused with the predictive value that *missing data itself* can offer (see Approaches to Data Missingness). For diagnostic records, differences in clinical inclusion and exclusion criteria between institutions can introduce label noise and give way to label bias.[34] Shifting criteria also lay the foundation for the difficulties in managing syndrome-based conditions, which were discussed in Challenges with syndromes. Temporal changes in data might occur as care and monitoring processes transform, including the disruption of existing clinical workflows resulting from implementing ML models. Taken together, these "systemic factors" all increase the heterogeneity of a clinical dataset which can confound predictive accuracy of a model not trained to recognize and correct for them (see Machine learning-based solutions, for an introduction to such methods). Regular updates to models are necessary as institutions evolve in these respects.

Differences in patient demographics between institutions can further add to the previously described data heterogeneity. However, this aspect is more difficult to handle due to the simultaneous potential for demographics to confound and improve impact predictive accuracy. Consider the following example, recent facial recognition[35] and hiring and recruitment[36] models were found to unintentionally perpetuate discriminatory harm as a result from overrepresentation of specific racial and ethnic groups during development and validation. Although the ML models themselves were independent of any explicit racial or ethnic bias, a biased data distribution contributed to the skewed outcomes[37]; this is one of many identified mechanisms behind "unfair" models.[34] In health care, large datasets may not be representative of traditionally underrepresented minorities,[38] which may similarly lead to bias. However, a demographic-specific biological susceptibility/response might also

contribute to unequal distributions of patients,[39,40] and this could help improve prediction. Recent data evaluating clinical use of ML models suggest that a significant number of them did not have any evaluation of racial bias,[41] as there is currently no standardized process to do so. Of the exceptions, studies designed to predict mortality or sepsis in critically ill patients were shown to be free of any bias.[42] Further investigation into this field is therefore indicated, and careful attention must be given to precisely delineate between the two effects and their degrees of impact on patient data.

Approaches to Data Missingness—A common theme to the obstacles of generalizability thus far is data heterogeneity.[8] Missingness similarly contributes another dimension to the data, increasing its heterogeneity. Several methods have been proposed and implemented in the clinical prediction models to handle this challenge, including omission, imputation, and physiology-focused solutions. These are applied depending on the type of missingness that is present in each problem. It should also be noted that existing studies on the clinical uses of ML do not agree on a universal practice, even for the same missingness mechanism. Our goal for this section is to therefore present the prominent techniques for handling missingness to provide clarity to the existing literature.

Of the methods that can handle missing data, omission is the simplest and least computationally demanding, an important consideration for processing large health care datasets.[43] The most reported process of omission for studies of prediction models that use ML is complete case analysis (CCA).[6] As its name suggests, CCA only includes patient cases that are not missing any data on the variables of interest.[44,45] However, its use is limited to datasets with MCAR data, as applying CCA to MAR or MNAR studies can introduce bias from nonrandom deletion of patients.[46–48] Consider a dataset with MAR data, younger diabetic patients have a higher rate of missing self-reported blood sugar data versus older diabetic patients as a result from only beginning the recording regimen. If CCA is applied here, we will disproportionately delete data from younger patients, so an ML model trained on this data will be biased toward making predictions for older diabetic patients. Nevertheless, even if CCA is correctly used for MCAR data, it still suffers from decreasing statistical power,[49] a flaw inherent to data omission.

More complex than methods of omission are those of imputation. Unlike omission, imputation requires added computation to produce more complete datasets, typically with less bias.[50] Three broad subtypes of imputation can achieve this with varying degrees of success: simple, hot deck, and multiple imputation.[45] Simple imputation replaces all missing values of a particular parameter with a single value computed from the present data. Hot deck imputation replaces a patient's missing value with one computed from a subset of patients with similar characteristics. This is repeated for each patient with missing data. Multiple imputation involves multiple rounds of simple imputation with minor changes; for a missing patient value, one is computed from a *random subset* of the existing dataset. Additional values are then computed from different random subsets, resulting in multiple possible replacement values. An aggregate of these values is taken (eg, mean), which ultimately replaces the missing one. Although imputation usually results in decreased bias of the dataset when compared with methods of omission on MAR or MNAR data, simple imputation may increase bias.[27] Specifically, for a large subset of missing data, imputation

with a single value can artificially decrease the variability of the dataset. Models trained on such data may subsequently have decreased generalizability to patient data from external cohorts. Hot deck and multiple imputations do not share this immediate weakness and generally result in more balanced datasets. Yet unlike simple imputation, they are more computationally expensive to perform.

Simple, hot deck, and multiple imputation methods provide a collective foundation for increasingly sophisticated variations of imputation. These should be thoughtfully used for predictive tasks because careless estimation of missing values can introduce artifacts that impact model accuracy. Specific factors that influence the type of imputation method used include study objectives, importance of missing data to these objectives, the amount of missing data, and the learning method used in an ML model.[51] As such, Table 2 summarizes major studies of ML use in clinical contexts and their imputation and learning methods.

We have thus far discussed statistical methods of handling classic mechanisms of missingness. Although these are useful for completing datasets with minimal bias,[44] predictive models may actually find additional physiologic value in the missing data itself, a process known as the missing indicator method (MIM).[55,56] A critical distinction must be made here, as omission and imputation methods are advantageous at completing datasets with minimal bias, they optimize for parameter estimation.[57] Alternatively, MIM may be more suitable for clinical prediction because it encompasses implicit factors that may be relevant to patient outcomes.[58,59] To clarify this concept, consider a dataset where all patients had a white blood count (WBC) ordered but at different times (4 AM and 4 PM). Agniel and colleagues demonstrated that patients with a normal 4 AM WBC were associated with a greater mortality rate than those with an abnormally high or low 4 PM WBC. Here, missingness of the 4 AM WBC is associated with an improved patient outcome, which is likely due to the fact that higher acuity patients require closer monitoring throughout the day and night.[16] As this example demonstrates, the *type* of missing data should be factored into its missingness pattern because the frequency of clinical and laboratory measurements are dictated by different guidelines and indications—not all missing data are of equal importance. Importantly, caution should be used with MIM approaches as this relies heavily on local practices and behaviors and this may lead to poor generalizability. Although MIM might introduce bias in estimation of causal relationships,[60] this can effectively improve predictive performance, a phenomenon termed “Stein’s Paradox.”[7,61]

When it comes to missingness, the prediction models benefit from statistical methods and MIM. MIM can provide additional insight into the state of the patient. However, it may decrease model generalizability because it strongly assumes that both the validation and development cohorts exhibit the same missingness mechanism. This is likely not the case, as missingness can be highly dependent on context. Over-reliance on MIM also limits the utility of a model as it considers information encoded in ordering clinical tests; they are consequently geared toward quantifying physician thinking instead of suggesting previously unconsidered diagnoses.[15] On the other hand, causal information in predictive models permits counterfactual prediction, a crucial component of models that provide the basis of a clinical decision.[62] How a prediction model handles missingness accordingly depends on its purpose. A focus on generalizability and counterfactual prediction should prioritize

statistical methods (ie, omission/imputation), and attention to a specific environment and risk assessment should prioritize MIM.[63,64] Being that each method has its own use case, many studied models use a combination of the two rather than complete reliance on either. Still, recent reviews of ML models have frequently demonstrated an insufficient or ignored method of handling missing data.[65–70]

MACHINE LEARNING-BASED SOLUTIONS

There exist various ML methods which may improve generalizability of predictive models for commonly encountered syndromic conditions in the ICU. They factor in the systemic differences between institutions described in Heterogeneity in Health Care. We focus on three approaches which critical care physicians may encounter when evaluating an ML model: transfer learning, continual learning, and conformal prediction. Each of these approaches has distinct methodologies and benefits which can yield improved performance under various situations.

Transfer Learning—Transfer learning is a technique in ML which has seen select use in health care. The current applications have been largely limited to oncology[71] and medical imaging,[72] with only recent applications into critical care. Conceptually, understanding transfer learning can be illustrated as follows: a prediction model (eg, to predict delayed septic shock) is developed and validated at a single institution. Although it can immediately be applied to a second institution, subtle differences between the locales (see Heterogeneity in Health Care) often dictate that the ML model's performance will be inferior to that of the original development institution. The initial development and validation of the ML model on a larger dataset may alleviate this drop in predictive ability, but this is not always possible. Transfer learning offers a solution to these subtle differences by using a small and representative dataset from the new location to optimize model parameters for it[73] (Fig. 1). Importantly, this bypasses the significant cost and data required for development and validation of a novel ML model per distinct institution to achieve similar performance. The utilization of a smaller dataset for retraining and fine-tuning of the original model is more computationally efficient and allows smaller hospitals to use such tools. However, being that the model must undergo retraining at each unique location, regulatory concerns arise as novel variants of the original model accumulate on a large scale. Nevertheless, examples of transfer learning applications in critical care include (1) fine-tuning a tracheal intubation prediction model for patients with COVID-19 pneumonia,[74] (2) adapting a delayed septic shock prediction model for use at various external institutions,[11] (3) predicting mortality in patients with end-stage renal disease,[75] (4) predicting acute kidney injury,[76] and (5) predicting acute respiratory distress syndrome on radiographs.[77] Test characteristics in these scenarios were significantly improved with the use of transfer learning.

Continual Learning—Continual learning (also referred to as lifelong learning, incremental learning, or sequential learning) describes a model that continuously learns and evolves based on increasing data, fed over time, with retention of previously gained knowledge.[78–81] In this way, it is intuitively appealing and like human cognition. One well-known use of continual learning is found in recommender systems of companies such as Amazon and Netflix—the model is continually updated with labeled data from interactions

with the end-user to reflect changes in personal preference over time.[82] Yet unlike transfer learning, research into its clinical applications has been meager in the critical care domain, and no implementations currently exist[83] because of the considerable potential for "catastrophic forgetting." This describes a phenomenon in which new information interferes with previously learned patterns, resulting in a paradoxic decrease in model performance. There also exist privacy concerns as the single model is continuously fed sensitive clinical data from various institutions. Observational data suggest this approach *may* improve predictive models over time, such as in the early prediction of sepsis,[84] medication dosing,[85] or augmentation of imaging studies performed in critically ill patients.[86] Nevertheless, these have not yet been translated into clinical tools.

Conformal Prediction—Conformal prediction refers to a model's assessment of the uncertainty of a prediction based on the past experience. Intuitively, when a model encounters a scenario like its training dataset, the confidence in prediction is high. However, when a model encounters a scenario where input data are significantly different (non-conformal) from training data, the utility and confidence of the prediction are uncertain. Conformal prediction is therefore a mathematical approach that quantifies the uncertainty of an ML prediction. In effect, this allows the model to say "I don't know" for inputs that are foreign from training data.[87–90] Applications of conformal prediction have been used in various nonmedical fields, including facial recognition, financial risk, and language recognition. In health care, it has been used to augment breast cancer diagnosis[91] and prediction assessment of stroke risk,[92] albeit primarily under research applications. Conformal prediction has more recently been described to assist in sepsis prediction, as shown in Fig. 2.[14] In this example, potential sepsis cases in which an ML model had low certainty in prediction were identified and resulted in a significant decrease in false alarms. The ideal application of conformal prediction in the ICU would follow a similar pattern: alerting clinicians to scenarios in which a predictive model has low certainty of a prediction may increase their trust in ML predictive scores.

DISCUSSION

ML models have the potential to significantly improve care of critically ill patients by leveraging the data-rich nature of the ICU. However, despite promising research, their real-world implementations in critical care are presently scant; this unfortunately growing chasm between what has been developed and what has been implemented is a phenomenon referred to as the "implementation gap."[93] Although there are various reasons behind this discrepancy, inherent challenges with generalizability in a syndrome-based, heterogeneous patient landscape have significantly limited the utility of ML models in critical care. Our review explores many of these obstacles. Various syndrome-based conditions with overlapping clinical characteristics are difficult for ML models to delineate due to a critical delay in patient information availability and ambiguity in syndrome diagnosis. These effects are especially pronounced in the ICU, where syndrome-based conditions are common, and patients rapidly evolve. Compounding the challenge is heterogeneity in health care data itself. Data recording and storage, local health care guidelines, and temporal shifts in data necessitate correction in models themselves. Although patient demographics and

missing data can further contribute heterogenous dimensions, they can also convey valuable information that might help improve prediction model accuracy.

Prevailing research on ML models use novel methods of learning to overcome these challenges and ultimately maximize external validity: transfer learning, continual learning, and conformal prediction.[11–14] Transfer learning involves retraining a model with limited data at a new deployment site so that it learns the specific nuances of that particular location.[73] Continual learning describes a central model that constantly adjusts its set of rules as it is applied to more data all while maintaining acceptable performance on prior applications.[79–81] Conformal prediction will only allow the model to predict from data that it deems "conformant" with the training data—that is, if it detects dissimilar data that may result in poor performance, it will refrain from making predictions.[87–90]

It is important to note that although we primarily demonstrate these in the context of sepsis prediction, ML models have been similarly studied for a variety of other prediction tasks, including respiratory failure in COVID-19 patients,[74,94–96] complications for critically ill patients,[76,77,97–99] and in-hospital mortality risk.[75,100–105] Other uses for clinical ML models outside of *predictive* tasks include identification of health factors related to patient outcomes, novel intervention design, and allocation of resources.[106] Future research into the applications of ML for critically ill patients should focus on *prospective* implementations across multiple centers to demonstrate clinical value. Many studies thus far are retrospective, and only a few of them undergo prospective validation; an even fewer number undertake randomized ML trials.[107] Indeed, many clinical ML models are currently developed and validated to collect dust only then in the "model graveyard."[93] Other deployed models, such as the Epic Sepsis Score (ESS), paint a cautionary tale to hospital systems that fail to perform rigorous testing and optimization during development: researchers at the University of Michigan described a significant performance drop and increased rate of false alarms of the ESS when implemented at their institution.[108] To conduct the multi-center trials necessary for demonstrating clinical value, models may therefore undergo local optimization. This can be accomplished via transfer learning or continual learning and further improved through conformal prediction or similar methods. Such approaches may help alleviate the problem of generalizability and improve test characteristics. To our knowledge, this has not yet been done and should thus be emphasized in future trial designs involving prospective validation of ML models.

SUMMARY

We believe that clinicians must understand the basics of ML and its major challenges to evaluate current and future models. This is a vital step for their successful implementation into clinical practice. Likewise, we believe that data scientists interested in the health care applications of ML must understand its unique clinical challenges; the barriers to generalizability can presently be overcome with solutions offered by transfer learning, continual learning, and conformal prediction. With increased attention, we are optimistic that a future with accurate and fair ML-based clinical aids is not far.

DISCLOSURE

G. Wardi is funded by the NIH (R35GM143121 and K23GM146092). A Malhotra is funded by the NIH. He reports income related to medical education from Livanova, Eli Lilly, Jazz. ResMed, Inc provided a philanthropic donation to UC San Diego in support of a sleep center. S. Nemati is funded by the National Institutes of Health (#R56LM013517, #R35GM143121, #R01EB031539) and the Gordon and Betty Moore Foundation (#GBMF9052). S. Nemati, S.P. Shashikumar, and A. Malhotra are cofounders and hold equity in Healcisio, although unrelated to this work. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict-of-interest policies. The remaining authors have no disclosures to report.

REFERENCES

1. Choy G, et al. Current applications and future impact of machine learning in radiology. Radiology 2018;288:318–28.
2. Chatterjee A, Somayaji NR, Kabakis IM. Abstract WMP16: artificial intelligence detection of cerebrovascular large vessel occlusion - nine month, 650 patient evaluation of the diagnostic accuracy and performance of the Viz.ai LVO algorithm. Stroke 2019;50. AWMP16-AWMP16.
3. van der Heijden AA, Abramoff MD, Verbraak F et al., Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System, Acta Ophthalmol, 96, 2018, 63–68.
4. Ratner M FDA backs clinician-free AI imaging diagnostic tools. Nat Biotechnol 2018;36:673–4.
5. Shashikumar S, Making AI algorithms safer. Nature Portfolio health community,Available at: http://healthcommunity.nature.com/posts/dddd. Accessed September 10, 2021.
6. Nijman S, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol 2022;142: 218–29.
7. van Smeden M, Groenwold RHH, Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. J Clin Epidemiol 2020;125:188–90. [PubMed: 32565213]
8. Luijken K, Groenwold RHH, Van Calster B, et al. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med 2019;38:3444–59.
9. Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit. Health 2020;2: e489–92. [PubMed: 32864600]
10. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.
11. Wardi G, Carlile M, Holder A et al. , Predicting progression to septic shock inthe emergency department using an externally generalizable machine-learning algorithm, Ann Emerg Med, 77, 2021, 395–406. [PubMed: 33455840]
12. Holder AL, Shashikumar SP, Wardi G, et al. A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the ICU. Crit Care Med 2021; 49:e1196. [PubMed: 34259450]
13. Amrollahi F, Shashikumar SP, Holder AL, et al. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. Sci Rep 2022;12:8380. [PubMed: 35590018]
14. Shashikumar SP, Wardi G, Malhotra A, et al. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. Npj Digit. Med. 2021;4:1–9.
15. Beaulieu-Jones B, Yuan W, Brat GA, et al., Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?, npj Digital Medicine, 4 (1), 2021, 62.
16. Agniel D, Kohane I, Weber G. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361:k1479. [PubMed: 29712648]
17. Brüggemann S, Chan T, Wardi G, et al. Decision support tool for hospital resource allocation during the COVID-19 pandemic. Inform Med Unlocked 2021;24:100618. [PubMed: 34095453]
18. Ye J, Yao L, Shen J, et al. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inf Decis Making 2020;20:295.

19. Hu C-A, Chen CM, Fang YC, et al. , Using a machine learning approach topredict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan, BMJ Open, 10 (2020), e033898.
20. Calvo F, Karras BT, Phillips R, et al. Diagnoses, syndromes, and diseases: a knowledge representation problem. AMIA Annu. Symp. Proc. AMIA Symp. 2003;802. [PubMed: 14728307]
21. Kawasaki T, Kosaki F, Okawa S, et al. A new infantile acute febrile mucocutaneous lymph node syndrome (MLNS) prevailing in Japan. Pediatrics 1974;54: 271–6. [PubMed: 4153258]
22. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016; 315:801–10. [PubMed: 26903338]
23. Hospital inpatient specifications manuals sepsis resources, Available at: https://qualitynet.cms.gov/inpatient/specifications-manuals/sepsis-resources. Accessed June 12, 2022.
24. Rhee C, Zhang Z, Kadri SS, et al., Sepsis surveillance using adult sepsisevents simplified eSOFA criteria versus sepsis-3 sequential organ failure assessment criteria, Crit Care Med, 47, 2019, 307–314. [PubMed: 30768498]
25. Seymour CW, Deutschman CS, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315:762–74. [PubMed: 26903335]
26. Rubin DB. Inference and missing data. Biometrika 1976;63:581–92.
27. Fielding S, Fayers PM, McDonald A, et al. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health Qual Life Outcome 2008;6:57.
28. Troxel AB, Fairclough DL, Curran D, et al. Statistical analysis of quality of life with missing data in cancer clinical trials. Stat Med 1998;17:653–66. [PubMed: 9549814]
29. Li C Little's test of missing completely at random. STATA J 2013;13:795–809.
30. Seaman S, Galati J, Jackson D, et al. What is meant by "missing at random". Stat Sci 2013;28:257–68.
31. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? Int J Epidemiol 2014;43:1336–9.
32. Rue T, Thompson HJ, Rivara FP, et al. Managing the common problem of missing data in trauma studies. J. Nurs. Scholarsh 2008;40:373–8. [PubMed: 19094153]
33. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255–60. [PubMed: 26185243]
34. Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169:866–72. [PubMed: 30508424]
35. Shellenbarger SA, Crucial step for averting AI disasters. WSJ, Available at: https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865. Accessed February 13, 2019.
36. Fawcett A, Understanding racial bias in machine learning algorithms. Educative: interactive Courses for Software Developers, Available at: https://www.educative.io/blog/racial-bias-machine-learning-algorithms. Accessed July 8, 2020.
37. Ferryman K, Pitcan M. Fairness in precision medicine. Data & society. Accessed February 26, 2018. Available at: https://datasociety.net/library/fairness-in-precision-medicine/.
38. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 2012;33:1–33.
39. Institute of Medicine (US). Committee on understanding and eliminating racial and ethnic disparities in health care. Unequal treatment: confronting racial and ethnic disparities in health care. National Academies Press: Washington DC (US); 2003.
40. Barnato AE, Alexander SL, Linde-Zwirble WT, et al. Racial variation in the incidence, care, and outcomes of severe sepsis. Am J Respir Crit Care Med 2008;177:279–84. [PubMed: 17975201]
41. Huang J, Galal G, Etemadi M, et al. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. JMIR Med. Inform. 2022;10: e36388. [PubMed: 35639450]
42. Allen A, Mataraso S, Siefkas A, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR Public Health Surveill 2020;6:e22400. [PubMed: 33090117]

43. Fang R, Pouyanfar S, Yang Y, et al. Computational health informatics in the big data age: a survey. ACM Comput Surv 2016;49. 12:1–12:36.
44. Donders ART, van der Heijden GJMG, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087–91. [PubMed: 16980149]
45. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons; 2019.
46. Buuren S van, Flexible imputation of missing data, Second Edition, 2018, CRC Press: Boca Raton, FL.
47. Harel O, Mitchell EM, Perkins NJ, et al. Multiple imputation for incomplete data in epidemiologic studies. Am J Epidemiol 2018;187:576–84. [PubMed: 29165547]
48. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338: b2393. [PubMed: 19564179]
49. Knol MJ, Janssen KJ, Donders AR, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010;63:728–36.
50. Liu D, Oberman HI, Muñoz J, et al. Quality control, data cleaning, imputation. 10.48550/ARXIV.2110.15877.
51. Syed M, Syed S, Sexton K, et al. Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: systematic review. Inform. MDPI 2021;8:16.
52. Davoodi R, Moradi MH. Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier. J. Biomed. Inform. 2018;79:48–59. [PubMed: 29471111]
53. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inf 2019;125:55–61.
54. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 2019; 321:2003–17. [PubMed: 31104070]
55. Sperrin M, Martin GP, Sisk R, et al. Missing data should be handled differently for prediction than for description or causal explanation. J Clin Epidemiol 2020;125:183–7. [PubMed: 32540389]
56. Galit Shmueli. To explain or to predict? Stat Sci 2010;25:289–310.
57. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 2007;60:979. [PubMed: 17689816]
58. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol 2019; 34:23–36. [PubMed: 30341708]
59. Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. J Mach Learn Res 2010;11:131–70.
60. Groenwold RHH, et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Can Med Assoc J 2012;184:1265. [PubMed: 22371511]
61. Efron B, Morris C. Stein's paradox stat. Sci Am 1977;236:119–27.
62. Lin L, Sperrin M, Jenkins DA, et al. A scoping review of causal methods enabling predictions under hypothetical interventions. Diagn. Progn. Res. 2021;5:3. [PubMed: 33536082]
63. Sisk R, Lin L, Sperrin M, et al., Informative presence and observation in routine health data: a review of methodology for clinical risk prediction, J Am Med Inform Assoc, 28, 2021, 155–166. [PubMed: 33164082]
64. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. Diagn. Progn. Res. 2020;4:8. [PubMed: 32699824]
65. Collins GS, Omar O, Shanyinde M, et al. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol 2013;66:268–77. [PubMed: 23116690]
66. Tsvetanova A, Sperrin M, Peek N, et al., Missing data was handled inconsistently in UK prediction models: a review of method used, J Clin Epidemiol, 140, 2021, 149–158.
67. Dhiman P, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol 2021;138:60–72. [PubMed: 34214626]

68. Galbete A, Tamayo I, Librero J, et al. Cardiovascular risk in patients with type 2 diabetes: a systematic review of prediction models. Diabetes Res Clin Pract 2022;184:109089. [PubMed: 34648890]
69. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol 2015;15:30. [PubMed: 25880850]
70. Karahalios A, Baglietto L, Carlin JB, et al. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodol 2012;12:96. [PubMed: 22784200]
71. Kim Y-G, et al. Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections. Sci Rep 2020;10: 21899. [PubMed: 33318495]
72. Alzubaidi L, Al-Amidie M, Al-Asadi A, et al. Novel transfer learning approach for medical imaging with limited labeled data. Cancers 2021;13:1590.
73. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172:1122–31.e9. [PubMed: 29474911]
74. Bendavid I, Statlender L, Shvartser L, et al. A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19. Sci Rep 2022;12:10573. [PubMed: 35732690]
75. Macias E, Morell A, Serrano J, et al. Mortality prediction enhancement in end-stage renal disease: a machine learning approach. Inform Med Unlocked 2020;19:100351.
76. Liu K, et al. Development and validation of a personalized model with transfer learning for acute kidney injury risk estimation using electronic health records. JAMA Netw Open 2022;5:e2219776. [PubMed: 35796212]
77. Sjoding MW, Taylor D, Motyka J, et al. Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. Lancet Digit. Health 2021;3:e340–8. [PubMed: 33893070]
78. Thrun S, Mitchell TM. Lifelong robot learning. Robot. Auton. Syst. 1995;15: 25–46.
79. Goodfellow IJ, Mirza M, Xiao D, et al., An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. 2013. doi: 10.48550/ARXIV.1312.6211.
80. Zenke F, Poole B & Ganguli S Continual Learning Through Synaptic Intelligence. in Proceedings of the 34th International Conference on Machine Learning 3987–3995 (PMLR, 2017).
81. van de Ven GM & Tolias AS Three scenarios for continual learning. (2019) doi:10.48550/arXiv.1904.07734.
82. Portugal I, Alencar P, Cowan D. The use of machine learning algorithms in recommender systems: A systematic review. Expert Syst Appl 2018;97:205–27.
83. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit. Health 2020;2:e279–81. [PubMed: 33328120]
84. French null. Catastrophic forgetting in connectionist networks. Trends Cogn. Sci. 1999;3:128–35.
85. Ghassemi MM, Alhanai T, Westover MB, et al., Personalized medication dosing using volatile data streams. In AAAI Workshops AAAI press, 2018, Available at: https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/17234.
86. Carlile M, et al. Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. J. Am. Coll. Emerg. Physicians Open 2020;1:1459–64.
87. Saunders C, Gammerman A, Vovk V. Transduction with confidence and credibility. Int Jt. Conf Artif. Intell. IJCAI 1999;16.
88. Vovk V, Gammerman A, & Saunders C (1999). Machine-Learning Applications of Algorithmic Randomness. International Conference on Machine Learning.
89. Papadopoulos H, Vovk V & Gammerman A Conformal prediction with neuralnetworks. in 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007) vol. 2 388–395 (2007).
90. Shafer G, Vovk V, A tutorial on conformal prediction. 2007. Available at: http://arxiv.org/abs/0706.3188. Accessed March 10, 2023.

91. Lambrou A, Papadopoulos H & Gammerman A Evolutionary Conformal Prediction for Breast Cancer Diagnosis. in 2009 9th International Conference on Information Technology and Applications in Biomedicine 1–4 (2009). doi:10.1109/ITAB.2009.5394447.
92. Papadopoulos H, Andreou A, Bramer M. Artificial intelligence applications and innovations. Springer: Larnaca, Cyprus.; 2010.
93. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innov 2020;6:45–7.
94. Bolourani S, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation (preprint). 2020. Available at: http://preprints.jmir.org/preprint/24246.
95. Ferrari D, Milic J, Tonelli R, et al. Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia—challenges, strengths, and opportunities in a global health emergency. PLoS One 2020;15:e0239172. [PubMed: 33180787]
96. Assaf D, Gutman Y, Neuman Y, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern. Emerg. Med. 2020;15: 1435–43. [PubMed: 32812204]
97. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019;23:112. [PubMed: 30961662]
98. Hyland SL, Faltys M, Hüser M., et al. , Early prediction of circulatory failure in the intensive care unit using machine learning, Nat. Med, 26, 2020, 364–373. [PubMed: 32152583]
99. Meyer A, Zverinski D, Pfahringer B, et al. , Machine learning for real-time prediction of complications in critical care: a retrospective study, Lancet Respir Med, 6, 2018, 905–914. [PubMed: 30274956]
100. Nanayakkara S, Fogarty S, Tremeer M, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: a retrospective international registry study. PLoS Med 2018;15:e1002709.
101. Di Castelnuovo A, Bonaccio M, Costanzo A, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutr. Metab. Cardiovasc. Dis. 2020;30:1899–913. [PubMed: 32912793]
102. Tezza F, Lorenzoni G, Azzolina D, et al. Predicting in-hospital mortality of patients with COVID-19 using machine learning techniques. J Pers Med 2021; 11:343. [PubMed: 33923332]
103. Du X, Min J, Shah CP, et al. Predicting in-hospital mortality of patients with febrile neutropenia using machine learning models. Int J Med Inf 2020;139:104140.
104. Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med Inf Decis Making 2020;20:251.
105. Brajer N, Cozzi B, Gao M, et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. JAMA Netw Open 2020;3:e1920733. [PubMed: 32031645]
106. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. Nat Mach Intell 2021;3:659–66.
107. Fleuren LM, Klausch TLT, Zwager CL, et al., Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy, Intensive Care Med, 46, 2020, 383–400. [PubMed: 31965266]
108. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181:1065–70. [PubMed: 34152373]

KEY POINTS

- Barriers to the use of machine learning in critically ill patients include challenges with recognition of syndromic conditions, data missingness, and the underlying heterogeneity of health care systems which may limit the generalizability of machine-learning algorithms.
- Recent advances in machine-learning applications, such as transfer learning and conformal prediction, can overcome barriers to improve generalizability across various institutions.
- Future studies are required to confirm the benefit of these strategies in both experimental and routine clinical care.

CLINICS CARE POINTS

- Syndromic conditions in the ICU are easy for clinicians to grasp, but many challenges exist for machine-learning models, which has thus far limited generalizability.
- Recent advances in machine-learning approaches may alleviate these concerns, although we still lack large, prospective trials demonstrating benefit in critically ill patients.

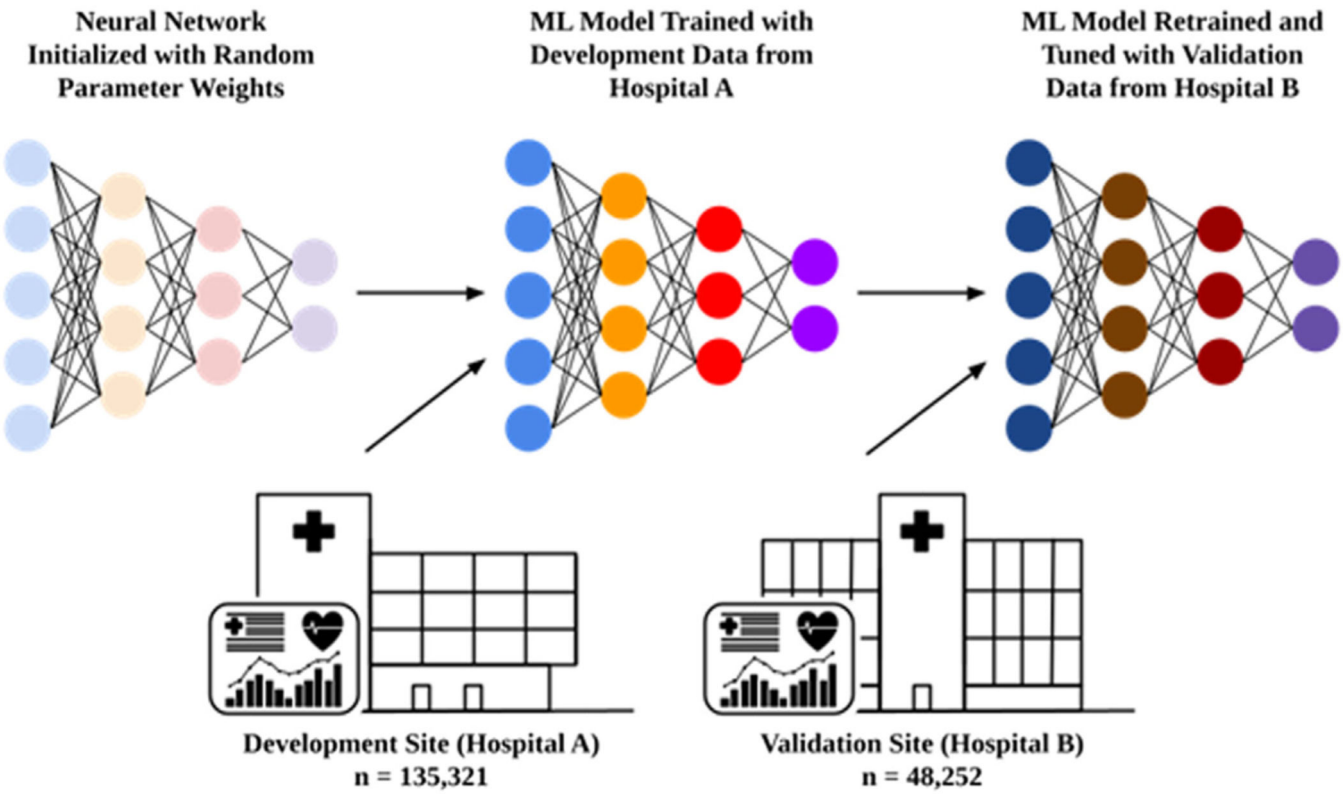

Fig. 1.

Example of applying transfer learning to a delayed septic shock prediction model. The initial ML model was fine-tuned using data at a second site. The use of transfer learning significantly increased test characteristics (AUCroc) of the delayed septic shock model at the validation site. AUCroc, area under the curve of the receiver operating characteristic curve. (*From* Wardi, G. et al Predicting Progression to Septic Shock in the Emergency Department Using an Externally Generalizable Machine-Learning Algorithm. Ann. Emerg. Med. 77, 395–406 (2021); with permission.)

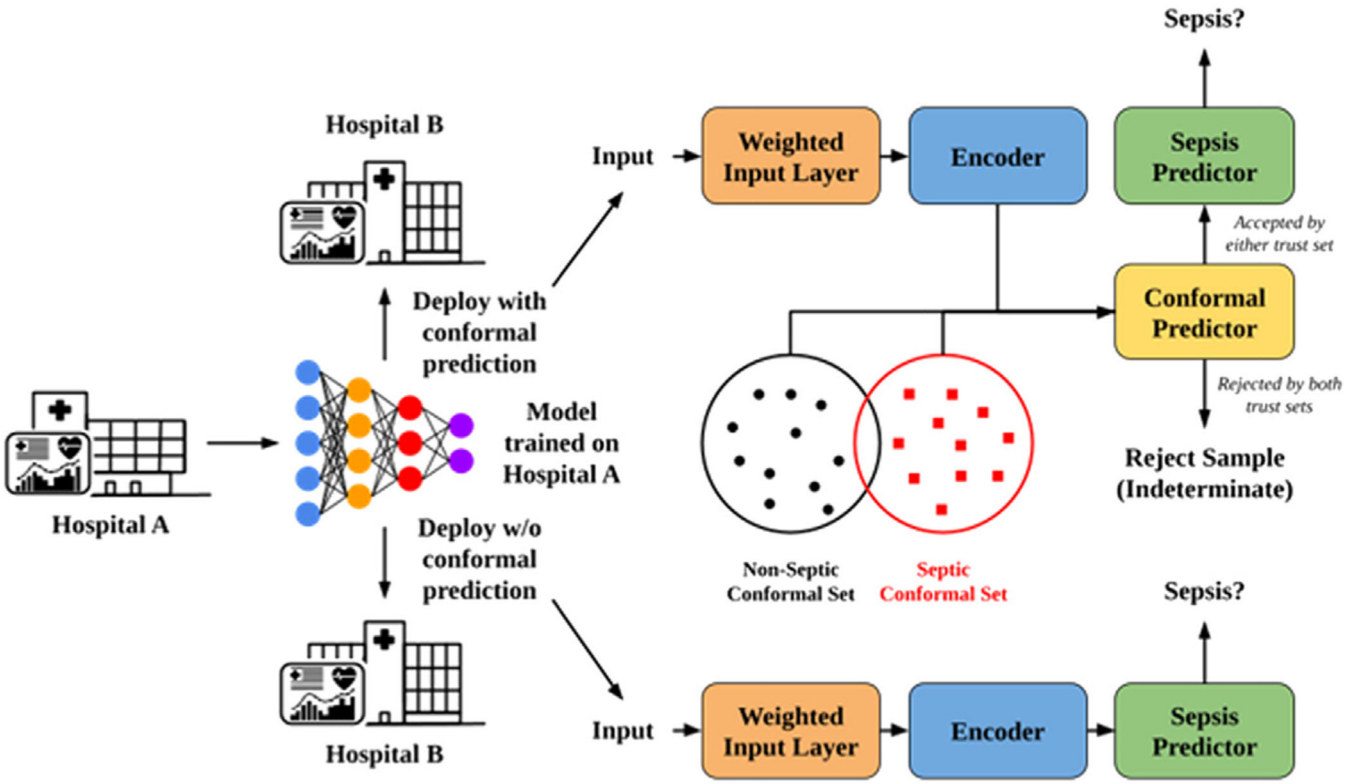

Fig. 2.

Example of applying conformal prediction to a sepsis prediction model. If the model does not recognize the input data from a patient, the conformal prediction layer "rejects" the data, and the sepsis prediction layer alerts the end-user that there exists a high degree of uncertainty. This resulted in a significant decrease in false alarms. (*From* Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". NPJ Digit Med. 2021;4(1):134. Published 2021 Sep 9. doi:10.1038/s41746–021-00504–6.)

Table 1

Definitions and examples of common terms used in data science

Term	Definition	Example
Development cohort	The dataset on which an ML model was trained	AISE[11,12] was trained on data from the Emory Healthcare System, also known as the development cohort.
Validation cohort	The dataset on which an ML model must perform a task for the first time	AISE[11,12] was validated on data from the MIMIC-III cohort collected from the Beth Israel Deaconess Medical Center in Boston, also known as the validation cohort.
Generalizability	The extent to which an ML model can achieve similar performance on external validation tasks vs internal validation tasks	AISE[11],[12] WUPERR,[13] and COMPOSER[14] demonstrated comparable performance at their respective validation hospitals, on par with their performance at their respective development hospitals.
Dimension	A characteristic of data and/or a dataset, relating to the number of variables included	EHR data are multidimensional because it includes heart rate, blood pressure, temperature, laboratory values, and so forth.
Missingness	The presence of missing clinical variables and how this is handled/reported	During Patient A's 10-d ICU stay, he/she may not have regular blood pressure measurements recorded every 4 h due to movements for scans, tests, and so forth.
Omission	A method of handling missingness where patients and/or characteristics with missing data are deleted from the dataset	As Patient A is missing blood pressure data, he/she is excluded from analysis of the dataset, which uses blood pressure.
Imputation	A method of handling missingness that determines a single or multiple value(s) to replace all missing values for a specific variable	Patient A's missing blood pressure measurements are estimated by averaging blood pressure values of other patients in the ICU taken at the same time (single imputation).

They are summarized here to provide clarity to the following sections.

Table 2

Major studies of clinical applications of machine learning, method of imputation for handling missingness and learning method used

Study	Missingness Method Used	Learning Method Used
Development and validation of an ICU mortality prediction model (Davoodi *et al*, 2018)[52]	Gaussian Imputation by Chained Equation	Deep rule-based fuzzy model
Development and validation of an in-hospital mortality prediction model for AKI (Lin *et al*, 2019)[53]	Mean Imputation	Random forest
Derivation and validation of novel sepsis phenotypes (Seymour *et al*, 2019)[54]	Multiple Imputation with Chained Equations	Latent class analysis
Development and validation of a volume responsiveness prediction model in oliguric AKI (Zhang *et al*, 2019)[97]	Multivariate Imputation by Chained Equation	XGBoost
Sepsis predictive model designed to identify instances of ML prediction uncertainty (Shashikumar, *et al* 2021)[14]	Mean Imputation and Sample-and-Hold with a weighted input layer to learn the "hold" duration	Conformal prediction

Abbreviation: AKI, acute kidney injury.