

Research and Applications

Reformulating patient stratification for targeting interventions by accounting for severity of downstream outcomes resulting from disease onset: a case study in sepsis

Fahad Kamran, PhD¹, Donna Tjandra , PhD¹, Thomas S. Valley, MD^{2,3}, Hallie C. Prescott, MD^{2,3}, Nigam H. Shah , MBBS, PhD⁴, Vincent X. Liu, MD⁵, Eric Horvitz , MD, PhD⁶, Jenna Wiens, PhD^{1,*}

¹Division of Computer Science and Engineering, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, United States, ²Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, United States, ³VA Center for Clinical Management Research, Ann Arbor, MI 48105, United States, ⁴Department of Medicine—Center for Biomedical Informatics Research, Clinical Excellence Research Center, Stanford University, Stanford, CA 94305, United States, ⁵Division of Research, Kaiser Permanente, Oakland, CA 94611, United States, ⁶Office of the Chief Scientific Officer, Microsoft, Redmond, WA 14820, United States

*Corresponding author: Jenna Wiens, PhD, Division of Computer Science and Engineering, Department of Electrical Engineering and Computer Science, University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, United States (wiensj@umich.edu)

Abstract

Objectives: To quantify differences between (1) stratifying patients by predicted disease onset risk alone and (2) stratifying by predicted disease onset risk and severity of downstream outcomes. We perform a case study of predicting sepsis.

Materials and Methods: We performed a retrospective analysis using observational data from Michigan Medicine at the University of Michigan (U-M) between 2016 and 2020 and the Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2012. We measured the correlation between the estimated sepsis risk and the estimated effect of sepsis on mortality using Spearman's correlation. We compared patients stratified by sepsis risk with patients stratified by sepsis risk and effect of sepsis on mortality.

Results: The U-M and BIDMC cohorts included 7282 and 5942 ICU visits; 7.9% and 8.1% developed sepsis, respectively. Among visits with sepsis, 21.9% and 26.3% experienced mortality at U-M and BIDMC. The effect of sepsis on mortality was weakly correlated with sepsis risk (U-M: 0.35 [95% CI: 0.33-0.37], BIDMC: 0.31 [95% CI: 0.28-0.34]). High-risk patients identified by both stratification approaches overlapped by 66.8% and 52.8% at U-M and BIDMC, respectively. Accounting for risk of mortality identified an older population (U-M: age = 66.0 [interquartile range—IQR: 55.0-74.0] vs age = 63.0 [IQR: 51.0-72.0], BIDMC: age = 74.0 [IQR: 61.0-83.0] vs age = 68.0 [IQR: 59.0-78.0]).

Discussion: Predictive models that guide selective interventions ignore the effect of disease on downstream outcomes. Reformulating patient stratification to account for the estimated effect of disease on downstream outcomes identifies a different population compared to stratification on disease risk alone.

Conclusion: Models that predict the risk of disease and ignore the effects of disease on downstream outcomes could be suboptimal for stratification.

Key words: patient stratification; downstream outcomes; causal effect estimation; heterogeneous effects.

Introduction

Predictive scoring methods are designed to help clinicians target treatments selectively with the ultimate goal of improving patient outcomes. To date, prevailing patient stratification efforts, whether relying on heuristic scores or machine learning models are aimed at identifying individuals at risk of developing disease. Such a stratification approach overlooks the heterogeneous effects of disease on patient outcomes.^{1–15} Our study addresses this oversight in current predictive scoring systems and presents steps forward for reformulating

patient stratification to consider both likelihood and severity of illness.

We hypothesize that the optimal way to define patient stratification will depend on the model's use case. In our example of selectively targeting prevention efforts for some disease due to resource constraints, our ultimate goal is to reduce patient mortality. A policy of triaging attention and resources by risk of developing disease makes an implicit assumption that the risk of the onset of a disease is representative of the risk of experiencing poor outcomes due to

developing disease. However, the validity of this assumption—and the implications of its potential invalidity for prioritizing interventions—have not been well studied.

Patients at low risk of developing disease may be likely to suffer death or complications should they develop disease. Under unavoidable constraints on both attentional and in-world resources, stratifying purely on the risk of disease may delay treatment for those who may be less likely to develop disease but more likely to suffer complications or die should they acquire the disease. In real-world clinical scenarios, triage typically considers the joint likelihood of the event and downstream consequences when allocating resources.^{16,17} For example, during the COVID-19 pandemic, vaccine allocation plans were not determined solely by an individual's propensity to get infected, but also by an individual's risk of COVID-19 complications and potential to spread to others.^{17–19} However, the importance of such a methodology has not been considered in predictive scoring. We note particularly that researchers pursuing AI in medicine have focused on leveraging machine learning for predicting the likelihood of illness, despite the relevance of severity of outcomes for clinical decision-making.²⁰

We study the importance of accounting for downstream consequences using a proof-of-concept study in the context of patient stratification tools for sepsis. Over the last decade, numerous patient stratification tools have been developed to help with early warning and response to the rise of sepsis.^{15,21–25} These tools consistently focus on the risk of developing sepsis^{15,21–25} and ignore disease severity as measured by the effect of sepsis on mortality. To probe the potential shortcomings of patient stratification based on likelihood of disease alone, we measured the heterogeneity in the effect of sepsis on risk of mortality within 2 large clinical cohorts and compared these results with the estimated risk of developing sepsis.

Materials and methods

Study cohort

We considered 2 retrospective cohorts extracted from electronic health record (EHR) datasets. The first included adults admitted to Michigan Medicine at the University of Michigan (U-M) between 2016 and 2020. In our primary analysis, we focused on only admissions to the intensive care unit (ICU). The second cohort included adults admitted to the ICU at Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2012.²⁶ Further inclusion and exclusion criteria can be found in the [Supplementary Material S1](#). In a secondary analysis involving the U-M dataset, we did not limit ourselves to ICU-only admissions and included admissions across the entire hospital (see [Supplementary Material S6.3](#)). However, our primary analysis only considered ICU encounters since the BIDMC dataset only contained ICU encounters. The use of the U-M dataset was approved by the institutional review board at U-M (HUM00176141). The BIDMC cohort is deidentified and available through Physionet.²⁶

Outcome definitions

In the U-M cohort, we defined sepsis using a composite definition based on meeting either (1) the clinical surveillance definition created by the Centers for Disease Control and Prevention (CDC) or (2) the Centers for Medicare and Medicaid Services (CMS) definition.^{27–30} Sepsis onset was defined as

the later time of when either the Systemic Inflammatory Response Syndrome (SIRS) criteria or the organ dysfunction criteria were met (for those meeting the CMS definition) or the first time in which the CDC definition was met (for those not meeting the CMS definition). For the BIDMC cohort, information necessary to define the CDC definition could not be obtained. Instead, in line with past work, we used a pragmatic definition based on the Sepsis-3 criteria, defining onset time by identifying the time of the acquisition of a body fluid culture temporally contiguous to the administration of antibiotics.^{15,31} For both cohorts, in-hospital mortality was identified from discharge information, where mortality was documented as indicated by the entries in the EHR.

Feature extraction

To build accurate effect estimates, we considered relevant features that may act as confounders between the development of sepsis and in-hospital mortality. To do so, for all patient admissions, we extracted demographics, vital sign measurements, laboratory test results, and nursing score information, including Glasgow coma scores and sedation information, throughout the hospitalization up until discharge or when the sepsis criteria were met. In the U-M cohort, we also considered vital signs and comorbidities from encounters within the past year for making predictions. Data were processed using FIDDLE with the default settings (see further details in the [Supplementary Material S2](#)).³²

Model development and evaluation: estimating sepsis risk and effect of sepsis on the risk of mortality

Developing sepsis has a direct effect on the risk of mortality, but this effect may be heterogeneous among patients. While past work has focused on estimating treatment benefits,^{33–35} we assumed a setting in which we aim to target some novel intervention not present in the available data (eg, additional monitoring) ([Figure 1](#)). We used machine learning and causal inference techniques to estimate an individual's risk of sepsis and the increase in the risk of in-hospital mortality if the individual were to develop sepsis. We split the data for each cohort into development and evaluation cohorts (for details see [Supplementary Material S1](#)). Our model development pipeline for estimating the risk of sepsis and the effect of sepsis on mortality is described below and summarized in [Figure 2](#).

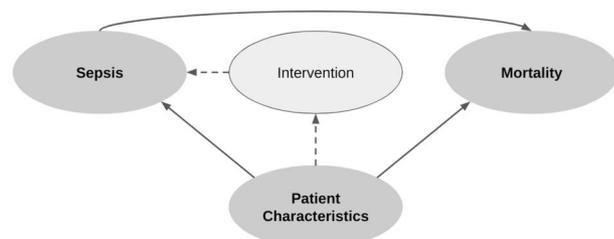


Figure 1. Assumed causal graph. The dashed lines represent causal relationships for the treatment that are not currently captured in the available data. Patient characteristics affect the risk of sepsis and mortality, all of which are fully observed in our data. Sepsis also affects the risk of mortality. Finally, there exists a potentially novel intervention currently not observed in the data. Our goal is to understand how to allocate interventions to patients to prevent sepsis and reduce the overall mortality rate.

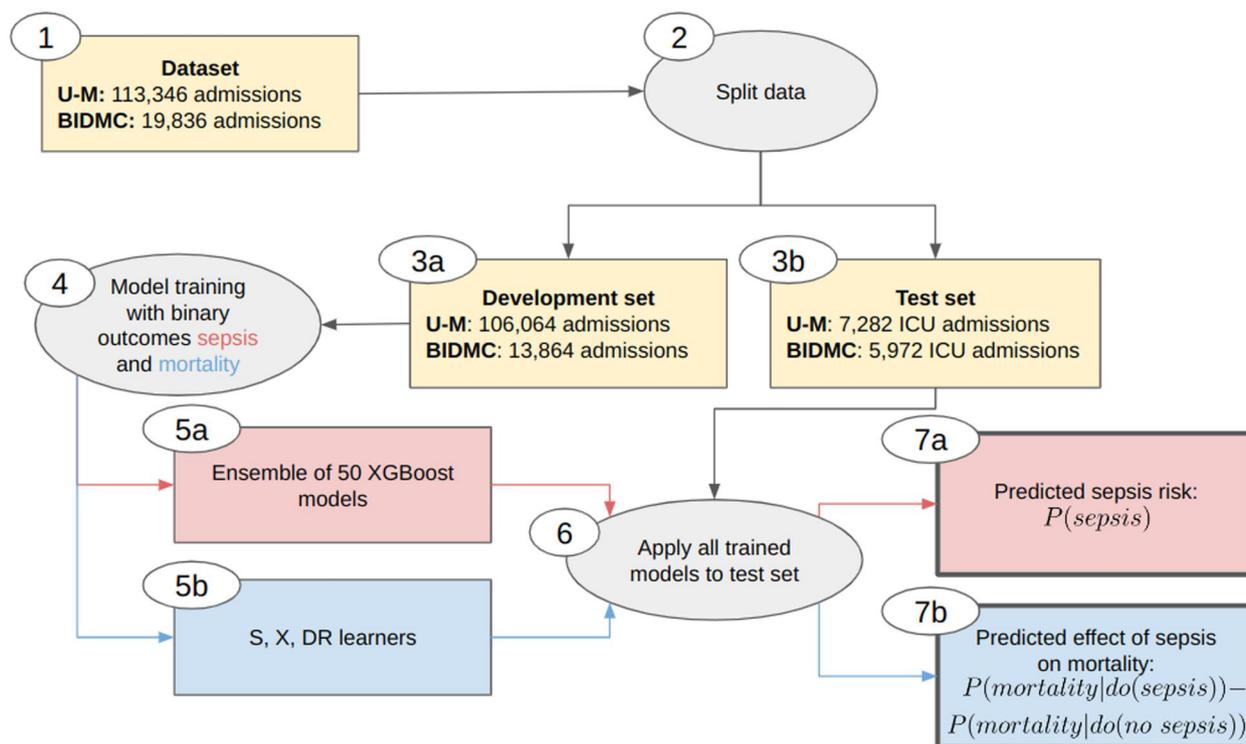


Figure 2. Summary of experimental setup. We begin with the entire dataset for both cohorts (1) and then split the dataset (2) into subsets for model development and test-time evaluation (3a, 3b). For the U-M dataset, we split temporally to simulate a setting in which a model is trained using retrospective data and is used prospectively. For the BIDMC dataset, we split based on a random shuffle since dates of service are obfuscated. Using the development set, we train (4) an ensemble of XGBoost models to predict the risk of sepsis (5a, shown in red), each trained with a randomly sampled 1-hour window from the admission. We train the S, X, and DR learners to predict the effect of sepsis on mortality (5b, shown in blue). Using these models, we obtain predictions on the test set (6) that are used for downstream analyses (7a, 7b).

Estimating the risk of sepsis (sepsis risk)

We build on prior work in machine learning for predicting the likelihood of sepsis.^{4,36–38} To estimate an individual’s risk of sepsis, based on the success of past work, we trained an ensemble of XGBoost models for each cohort to predict the risk of sepsis at every hour (for details see [Supplementary Material S3](#)). Applied to held-out evaluation cohorts, we evaluated the sepsis risk models in terms of the area under the receiver-operator characteristic curve (AUROC) at the hospital admission level.^{4,22,36} To calculate the AUROC at test time, we used the maximum predicted risk score within the admission as the model’s prediction. We used the maximum score since this mimics a use case in which a model flags patients as high-risk once they exceed some prespecified threshold; by taking the maximum we can sweep this threshold and calculate an ROC curve. We estimated the 95% CIs with 500 bootstrapped samples.

Estimating the effect of sepsis on the risk of mortality

Given the extracted confounders, to estimate the effect of sepsis on in-hospital mortality risk, we used 3 tools for computing conditional average treatment effects that adjust for confounding: (1) the S-Learner, (2) the X-Learner, and (3) the DR-Learner.³⁹ We applied these techniques independently to both cohorts (for training details see [Supplementary Material S4](#)). To validate each approach, we first evaluated the models’ ability to accurately predict mortality within both the population with sepsis and the population without sepsis in terms of the AUROC. We also performed a global

null analysis, separately retraining causal models using random “treatment” assignments in both the septic and nonseptic groups.^{40,41} On held-out evaluation cohorts, we checked whether or not the models could recover a global null treatment effect (ie, mean squared error between estimated effect and 0 was zero). We estimated 95% CIs with 500 bootstrapped samples. In line with best practice, we ran all analyses using all 3 approaches, checking for consistency.^{39,41} We report results of the S-Learner in the main paper and include remaining results in the [Supplementary Material S6.2](#).

Statistical analysis

Heterogeneity in the effect of sepsis on mortality

To inspect the effect of sepsis on in-hospital mortality risk at test time, we calculated the mean estimate for each admission across all windows. Here, and in subsequent analyses, we used the mean prediction instead of the maximum in order to remove bias due to longer admissions and to make the predictions between stratification schemes comparable since the time at which the maximum score for sepsis risk occurred did not always match the time of the maximum predicted effect of sepsis on mortality. We produced visualizations of the distribution of these estimated effects and reported the median of each evaluation cohort. To quantify heterogeneity, we calculated the difference between the 90th and 10th percentile of estimated effect sizes.

Correlation between risk of having sepsis and its effect on mortality

We calculated Spearman’s correlation between risk of sepsis and effect of sepsis on mortality. To estimate the relationship

between these variables at a per-admission level at test time, we aggregated all predictions by calculating the mean estimate for each admission across all windows. We estimated 95% CIs with 500 bootstrapped samples.

We visualized the relationship between the risk of developing sepsis and the effect of sepsis on mortality by plotting a random sample of patient data points as well as the mean and 95% CI of estimated effect of sepsis on mortality for each quintile of estimated risk of sepsis. We also visualized the empirical distributions of estimated effects for high-risk (ie, top 20%) and low-risk (ie, not top 20%) sepsis windows, respectively, where 20% was chosen to match the alert rate of existing sepsis risk stratification models.²²

Differences in stratified patient populations

Finally, we compared the cohort of individuals selected by a model focused only on the risk of sepsis to one selected by weighing *both* risk of sepsis and estimated effect of sepsis (see [Supplementary Material S6.2](#) for details). We stratified 20% of the patient population for intervention by either: (1) triaging individuals with the highest estimated risk of sepsis during their admission (Risk) or (2) triaging individuals based on the product of sepsis risk and estimated effect of sepsis on mortality risk during admission (Joint). We compared the 2 triaged cohorts based on demographic information to understand what cohorts of individuals may be deemphasized if the effects of sepsis on mortality risk are unaccounted for. To further understand the differences in these populations, we also examined the mortality rate and incidence of comorbidities within quintiles stratified by risk of sepsis ([Supplementary Material S6.4](#)).

Results

Model development sets included 106 064 U-M patient admissions (median age 57.0 years [interquartile range—IQR 37.0-69.0]) and 13 864 BIDMC patient admissions (median age 65.7 years [IQR 53.2-78.1]) ([Table S1](#)). Of these admissions, 5391 (5.1%) developed sepsis and 2014 (1.9%) experienced in-hospital mortality in the U-M development cohort, while 1108 (8.0%) developed sepsis and 1231 (8.9%) experienced in-hospital mortality in the BIDMC development cohort.

Our final evaluation cohorts consisted of 7282 and 5942 ICU stays in the U-M and BIDMC cohorts, respectively ([Table 1](#)). In the U-M evaluation cohort, 576 (7.9%) admissions developed sepsis and 574 (7.9%) admissions experienced in-hospital mortality. Within U-M, mortality rates were 21.9% and 6.7% for the septic and nonseptic populations, respectively. In the BIDMC evaluation cohort, 483 (8.1%) admissions developed sepsis, while 512 (8.6%) experienced in-hospital mortality. Within the BIDMC, in-hospital mortality rates were 26.3% and 7.1% in the septic and nonseptic populations, respectively.

For the task of predicting sepsis, our learned models achieved an AUROC of 0.69 (95% CI, 0.67-0.71) in the U-M cohort and 0.74 (95% CI, 0.72-0.77) in the BIDMC cohort. For the task of predicting mortality without sepsis and with sepsis, the S-Learner achieved AUROCs of 0.89 (95% CI, 0.87-0.90) and 0.79 (95% CI, 0.74-0.83), respectively, in the U-M cohort and 0.87 (95% CI, 0.85-0.88) and 0.77 (95% CI, 0.73-0.82), respectively, in the BIDMC cohort. The global null test showed that all models can accurately predict null treatment effects, with the S-Learner performing best ([Table S2](#)).

The histograms of estimated effect of sepsis on mortality risk confirms that the downstream effect is both positive and heterogeneous in both datasets ([Figure 3](#)). The S-Learner estimated a median effect of sepsis on mortality of 6.19 percentage points (pp) and 8.82pp in the U-M and BIDMC cohorts, respectively. The spread of estimated effect of sepsis on mortality between the 90th and 10th percentile was 15.4pp in the U-M cohort and 15.3pp in the BIDMC cohort. The X-Learner and DR-Learner led to similar results (see [Supplementary Material S6.2](#)).

Risk of sepsis and estimated effect of sepsis on mortality were weakly correlated in both datasets (U-M: 0.35 [95% CI, 0.33-0.37] and BIDMC: 0.31 [95% CI, 0.28-0.34]). Within quintiles of sepsis risk, there is large variability in effect of sepsis on mortality ([Figure 4A and B](#)). Among patient windows at or above the 80th percentile of sepsis risk, sepsis had only a small estimated effect on the increased risk of mortality (ie, <5pp) for 34.8% and 17.9% of patient windows in the U-M and BIDMC cohorts, respectively ([Figure 4C and D](#)). Meanwhile, for the remaining 80% of patient windows at lower risk of sepsis, developing sepsis was estimated to have a substantial increase in mortality risk (>20pp) for over

Table 1. Evaluation cohort characteristics.^a

	U-M evaluation cohort (n = 7282)	BIDMC evaluation cohort (n = 5942)
Female (%)	3242 (44.5%)	2603 (43.8%)
Median (IQR) age (years)	62.0 (48.0-72.0)	65.3 (53.0, 77.7)
Race		
White	5942 (81.6%)	4296 (72.3%)
Black	767 (10.5%)	645 (10.9%)
Asian	172 (2.4%)	164 (2.8%)
Other or unknown	401 (5.5%)	837 (14.1%)
Ethnicity		
Hispanic or Latino	176 (2.4%)	255 (4.3%)
Not Hispanic or Latino	6880 (94.5%)	5138 (86.5%)
Other or unknown	226 (3.1%)	549 (9.2%)
No. sepsis (%)	576 (7.9%)	483 (8.1%)
No. in-hospital mortality (%)	574 (7.9%)	512 (8.6%)
Within septic group (%)	126 (21.9%)	127 (26.3%)
Within nonseptic group (%)	448 (6.7%)	385 (7.1%)

^a Characteristics of the evaluation cohorts for both datasets. Characteristics of the development cohort can be found in the [Supplementary Material S6](#).

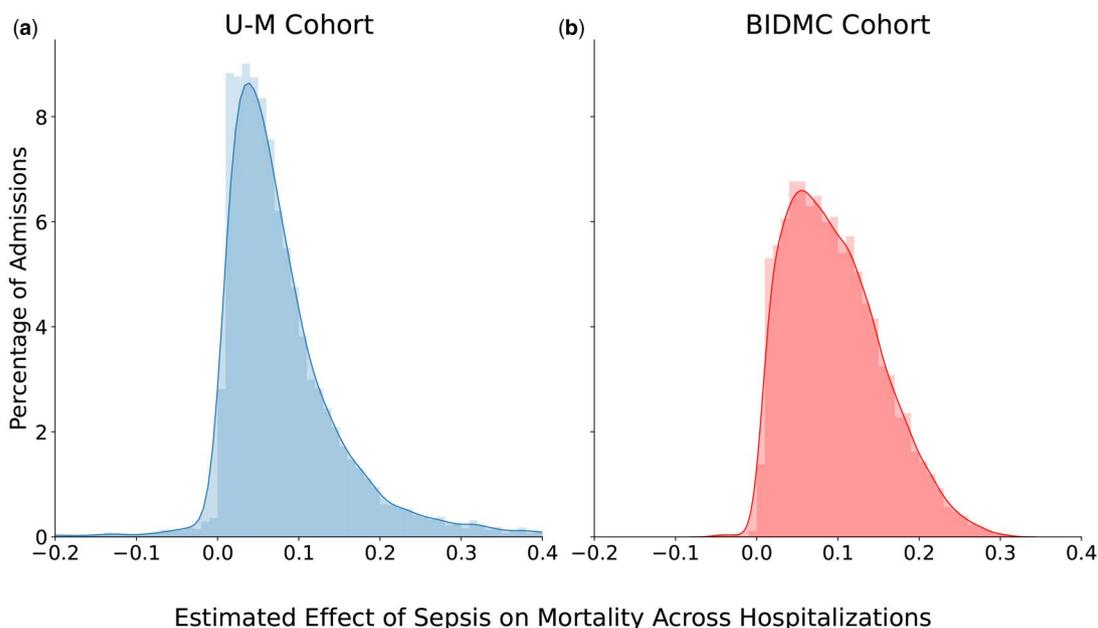


Figure 3. Estimated effect of sepsis on mortality. Estimated effect of sepsis on the risk of mortality across hospital admissions as estimated by the S-Learner. The average estimated effect is positive in (a) the U-M cohort and (b) the BIDMC cohort. Moreover, there is substantial heterogeneity in the estimated effect of sepsis on mortality.

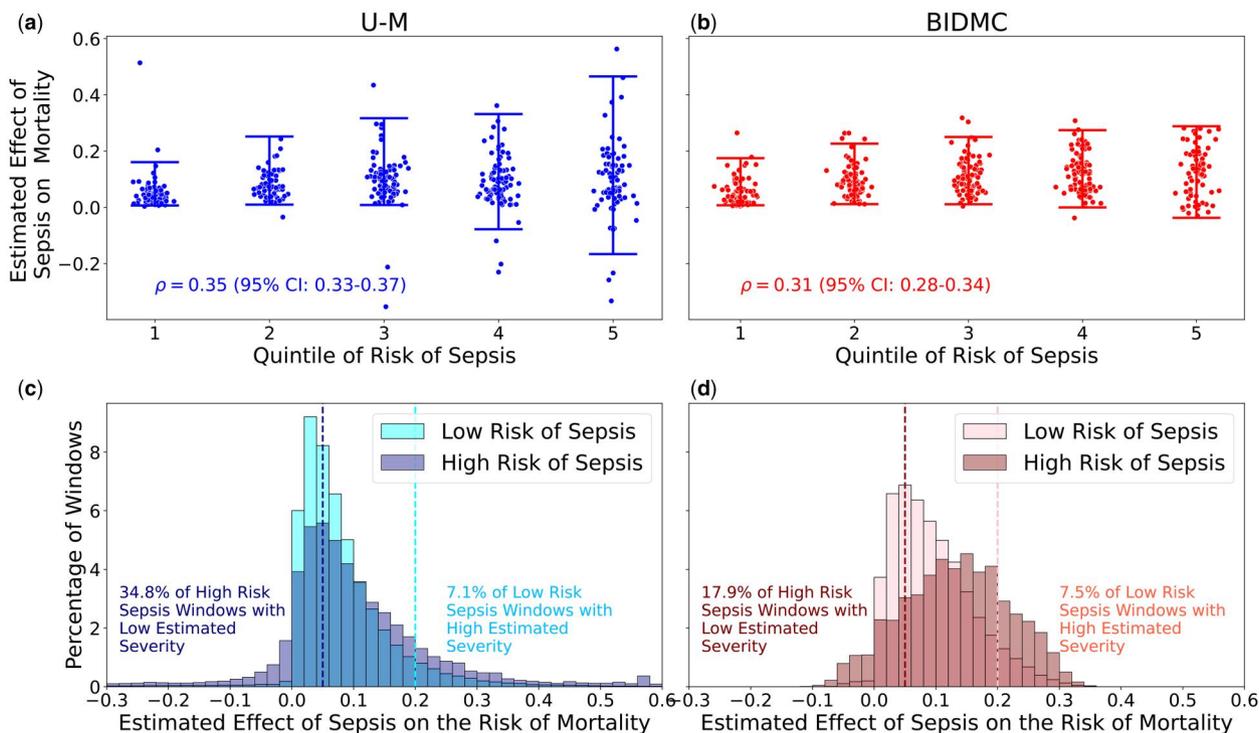


Figure 4. Relationship between the risk of having sepsis and the effect of sepsis on mortality, as estimated by the S-Learner. The estimated effect of sepsis on mortality is larger for patients within higher quintiles of risk of sepsis [top: (a) U-M cohort, (b) BIDMC cohort]. Meanwhile, there are many patients at high risk of sepsis who are still estimated to have a low effect of sepsis on their mortality, as well as many patients who are low risk for sepsis but would be severely adversely affected from developing sepsis [bottom: (c) U-M cohort, (d) BIDMC cohort].

7.1% of patient windows for each cohort. These overall findings hold for the other causal inference techniques, with most models showing a weak correlation between risk of sepsis and effect of sepsis on mortality (see [Supplementary Material S6.2](#)). These results also hold on the entire inpatient U-M cohort (see Figures S4 and S5).

The top 20% of admissions prioritized by the Risk and Joint triage approaches overlapped by 66.8% and 52.8% in the U-M and BIDMC cohorts, respectively. The demographics of the cohorts prioritized by each triage approach differed along several dimensions ([Table 2](#)). Consistently, the cohort of individuals identified by the Joint model was

Table 2. Triaged cohorts characteristics.^a

	U-M (<i>n</i> = 1457)		<i>P</i>	BIDMC (<i>n</i> = 1189)		<i>P</i>
	Triaged by Risk model	Triaged by Joint model		Triaged by Risk model	Triaged by Joint model	
Female (%)	461 (38.8%)	527 (44.3%)	.006	553 (38.0%)	562 (38.6%)	>.05
Median (IQR) age (years)	63.0 (51.0, 72.0)	66.0 (55.0, 74.0)	<.001	68.0 (59.0, 78.0)	74.0 (61.0, 83.0)	<.001
Race						
White	1116 (76.7%)	1160 (78.9%)	>.05	867 (72.9%)	824 (69.3%)	.046
Black	220 (15.1%)	186 (12.8%)	>.05	83 (7.0%)	101 (8.5%)	>.05
Asian	25 (1.7%)	32 (2.2%)	>.05	26 (2.2%)	41 (3.4%)	.039
Other or unknown	96 (6.6%)	79 (5.4%)	>.05	213 (17.9%)	223 (18.8%)	>.05
Ethnicity						
Hispanic or Latino	28 (1.9%)	21 (1.4%)	>.05	41 (3.4%)	33 (2.8%)	>.05
Not Hispanic or Latino	1371 (94.1%)	1367 (93.9%)	>.05	976 (82.1%)	966 (81.2%)	>.05
Other or unknown	58 (4.0%)	69 (4.7%)	>.05	172 (14.5%)	190 (16.0%)	>.05

^a Characteristics of the triaged cohorts for both datasets using different triaging models.

significantly older than that identified by the Risk model, as demonstrated by the difference in median patient age within each cohort (U-M: 66.0 years [IQR: 51.0-72.0] vs 63.0 years [IQR: 55.0-74.0], BIDMC: 74.0 years [IQR: 61.0-83.0] vs 68.0 years [IQR: 59.0-78.0], $P < .05$). This trend held across almost all causal inference techniques across both cohorts (Supplementary Material S6.2). In our analysis on the mortality rate and comorbidity incidence within quintiles stratified by risk of sepsis, we found that the mortality rates and comorbidity incidences tended to be higher at the upper quintiles of sepsis risk (Supplementary Material S6.4). This can help explain why the effect of sepsis on mortality in this group is smaller, as this may represent a population that is sicker at baseline.

Discussion

Standard patient stratification identifies patients at greatest risk of developing a condition. Using such a stratification to then target additional resources or interventions often introduces an implicit erroneous assumption: prioritizing those most at risk of developing the condition is an ideal policy for allocating interventions that aim to reduce downstream morbidity and mortality. However, beyond risk of developing a condition (such as sepsis), benefit from treatment may also depend on the effects of developing the condition on downstream outcomes (such as mortality).⁴² To surface a need to reformulate patient stratification in many areas of medicine, we undertook a proof-of-concept study for patient stratification for sepsis.

Across 2 clinical cohorts from different time periods and different hospitals, we found that the effect of developing sepsis on risk of in-hospital mortality was heterogeneous. Moreover, we consistently found that the risk of sepsis was not highly correlated with the effect of sepsis on mortality risk. These findings point to a limitation in standard patient stratification.

Measuring heterogeneity response to disease, we found that those most likely to develop sepsis may not always be more likely to experience poor outcomes due to it. Vice versa, many patients who would have the greatest increase in risk of mortality if they were to develop sepsis are not those most likely to develop sepsis. *Allocating interventions to the*

former rather than the latter could delay interventions to those who would most benefit. We found that age consistently remained an important source of heterogeneity. Standard patient stratification approaches may miss triaging older individuals who are adversely affected by the development of sepsis.

Although our analysis uses sepsis as a case study, our findings have broad implications when developing and evaluating patient stratification models for allocating interventions. For example, predictive models are being used to estimate the likelihood of numerous diseases in the hospital.^{4,43} The development and use of these models follow the same principle of allocating treatments to those most at risk of developing disease. However, the effect of developing disease on downstream complications, such as mortality, may be heterogeneous. For example, the clinical presentation of COVID-19 is heterogeneous, and preventative efforts should account for those who are most at risk of deterioration due to infection.⁴⁴ Our work complements past attempts to identify these severe cases of the disease by reframing the problem as *a combination of the risk of acquiring a condition and the individual effect of that condition on adverse patient outcomes*, with the goal of understanding how the disease may affect the likelihood of mortality.⁴⁵

Recent work has estimated heterogeneous treatment effects for optimal intervention allocation but does not consider intermediate outcomes, instead focusing on settings where the treatment directly prevents the final outcome.^{34,35,46} In contrast, we consider a setting in which an intermediate outcome (eg, sepsis onset) can occur, and the treatment prevents the final outcome (eg, mortality) indirectly by preventing the intermediate outcome. As a result, it becomes important to account for the heterogeneous effects of the intermediate outcome on the final/downstream outcome, as we have highlighted in this work. Doing so is especially important when there exists an intervention that could prevent the intermediate outcome, but the data needed to estimate its effect on the downstream outcome are unavailable. For example, during the COVID-19 pandemic, the implementation of preventative efforts, such as novel vaccines, focused on addressing both the likelihood of acquiring the virus (ie, the intermediate outcome) and the potential for severe complications (ie, the downstream outcome).¹⁷⁻¹⁹ Alternatively, in settings where

the treatment prevents the downstream outcome directly, our work also applies if the treatment is allocated proactively (ie, before the intermediate outcome) (see [Supplementary Material S5](#)). For example, in the context of COVID-19, incorporating the effect of acquiring the virus on mortality when determining the proactive allocation of Paxlovid⁴⁷ helped to reduce the time between a positive COVID-19 test result and treatment, increasing the effectiveness of the drug.

We note limitations in our study. First, causal inference techniques rely on untestable assumptions. Violating these assumptions may result in biased effect estimates. Moreover, due to the lack of ground-truth effects, we are unable to validate the learned effects of sepsis on mortality risk. To overcome this in part, in line with past work, we confirmed that our key takeaways held across a multitude of different causal inference techniques.^{39,41} Second, we assumed a particular causal model of the world ([Figure 1](#)). We stress that this model of the world is an oversimplification of the truly complex nature of sepsis and most diseases. However, we emphasize that this work is not meant to guide clinical practice in its current state, but rather, is a proof-of-concept study to demonstrate the importance of incorporating downstream patient outcomes into patient stratification tools. Moreover, when estimating the effect of sepsis on mortality, we are estimating the total effect of sepsis on mortality. This includes mediator variables that are treatments and represent the standard of care during the hospitalization, including antibiotics for patients who developed sepsis. However, in this study, as we assume that individuals with similar characteristics are given similar levels of treatment, heterogeneity due to treatments present in the data should not confound results. Additionally, for pragmatic reasons, we used different definitions of sepsis in each dataset. Despite this, our results are consistent across datasets. Finally, in our main analysis, we focused on ICU encounters due to limitations of the BIDMC dataset. Although ICU encounters are associated with greater mortality, the trends hold across a broader cohort of inpatient encounters at U-M (see [Supplementary Material S6.3](#)).

Overall, this study has significant implications for using predictive models to guide the allocation of attention and interventions. Our goal is to raise awareness and frame research on patient risk stratification across health-care disciplines rather than to provide specific guidance to clinical practice. In practice, the optimal way to stratify patients will depend on the model's use case itself. For example, a model that is intended to be used for hypothesis generation about the underlying disease process is likely to be effective when stratifying on risk of disease only. In our study, we examined a different case where the model's predictions are being used to guide resource allocation. We highlighted how the model's predictions (ie, risk of sepsis) did not directly align with the overall task (ie, resource allocation to reduce mortality). More generally, a model may have more than 1 use case, so one could potentially perform patient stratification in different ways during evaluation to provide a more comprehensive analysis that aligns with different clinical needs.⁴⁸ Nonetheless, our findings highlight the important limitations of existing methods that ignore the potentially heterogeneous effects that the acquisition of a condition may have on downstream patient outcomes. All agree the goal is to optimize patient outcomes. However, targeting interventions by risk of condition alone will miss individuals who are at lower risk but

more likely to experience poor outcomes, should the condition be acquired.

Author contributions

Fahad Kamran (Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization), Donna Tjandra (Data curation, Software, Validation), Thomas S. Valley (Methodology, Validation), Hallie Prescott (Methodology, Validation), Nigam H. Shah (Methodology, Validation), Vincent Liu (Methodology, Validation), Eric Horvitz (Conceptualization, Methodology, Supervision, Validation), and Jenna Wiens (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision)

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by Cisco Research, the National Institutes of Health (NIH), NIH grant number R35GM128672, and the National Science Foundation (NSF), NSF grant number IIS 2124127. The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Cisco Research or the NSF. This material is also the result of work supported with resources and use of facilities at the Ann Arbor Veterans Affairs (VA) Medical Center. This manuscript does not represent the views of the Department of Veterans Affairs or the US government.

Conflicts of interest

The authors have no conflicts of interest to report.

Data availability

The data underlying this article for the BIDMC cohort are available through Physionet at <https://physionet.org/content/mimiciii/1.4/>. The data underlying this article for the Michigan Medicine cohort cannot be shared due to current policies.

References

1. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)*. 2014;33:1123-1131. <https://doi.org/10.1377/hlthaff.2014.0041>
2. Wiens J, Gutttag JV, Horvitz E. Patient risk stratification for hospital-associated *C. diff* as a time-series classification task. *Adv Neural Inf Process Syst*. 2012;25:467-475.
3. Wiens J, Gutttag J, Horvitz E. Patient risk stratification with time-varying parameters: a multitask learning approach. *J Mach Learn Res*. 2016;17:1-23.
4. 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.
5. Puopolo KM, Draper D, Wi S, et al. Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. *Pediatrics*. 2011;128:e1155-1163. <https://doi.org/10.1542/peds.2010-3464>
 6. Goldstein BA, Pencina MJ. Developing implementable risk prediction models with electronic health records data. In: *Wiley StatsRef: Statistics Reference Online*. Published online 2014:1-8.
 7. Mani S, Ozdas A, Aliferis C, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J Am Med Inform Assoc*. 2014;21:326-336.
 8. Weiss J, Kuusisto F, Boyd K, Liu J, Page D. Machine learning for treatment assignment: improving individualized risk attribution. In: *AMIA Annual Symposium Proceedings*. Vol. 2015. American Medical Informatics Association; 2015:1306.
 9. Lee DS, Ezekowitz JA. Risk stratification in acute heart failure. *Can J Cardiol*. 2014;30:312-319. <https://doi.org/10.1016/j.cjca.2014.01.001>
 10. Olsen MA, Higham-Kessler J, Yokoe DS, et al. Developing a risk stratification model for surgical site infection after abdominal hysterectomy. *Infect Control Hosp Epidemiol*. 2009;30:1077-1083. <https://doi.org/10.1086/606166>
 11. Yang Y, Yang KS, Hsann YM, Lim V, Ong BC. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *J Crit Care*. 2010;25:398-405. <https://doi.org/10.1016/j.jcrrc.2009.09.001>
 12. Winters BD, Eberlein M, Leung J, Needham DM, Pronovost PJ, Sevransky JE. Long-term mortality and quality of life in sepsis: a systematic review. Published online. 2010. <https://doi.org/10.1097/CCM.0b013e3181d8cc1d>
 13. Leigdowicz A, Matthey MA. Heterogeneity in sepsis: new biological evidence with clinical applications. *Crit Care*. 2019;23:80-88. <https://doi.org/10.1186/S13054-019-2372-2/FIGURES/2>
 14. Ibrahim ZM, Wu H, Hamoud A, Stappen L, Dobson RJB, Agaroosi A. On classifying sepsis heterogeneity in the ICU: insight using machine learning. *J Am Med Inform Assoc*. 2020;27:437-443. <https://doi.org/10.1093/JAMIA/OCZ211>
 15. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. *Proc Mach Learn Res*. 2019;106:1. <https://github.com/BorgwardtLab/mgp-tcn>
 16. Christ M, Grossmann F, Winter D, Bingisser R, Platz E. Medicine modern triage in the Emergency Department. <https://doi.org/10.3238/arztebl.2010.0892>
 17. Kahn B, Brown L, Foege W, et al. A framework for equitable allocation of COVID-19 vaccine. *Framework for Equitable Allocation of COVID-19 Vaccine*. Published online 2020.
 18. Swift MD, Sampathkumar P, Breecher LE, Ting HH, Virk A. Mayo Clinic's multidisciplinary approach to Covid-19 vaccine allocation and distribution. *NEJM Catal Innov Care Deliv*. 2021;2. <https://doi.org/10.1056/CAT.20.0696>
 19. Dooling K, Marin M, Wallace M, et al. The Advisory Committee on Immunization Practices' updated interim recommendation for allocation of COVID-19 vaccine—United States, December 2020. *MMWR Morb Mortal Wkly Rep*. 2021;69:1657-1660.
 20. Prescott HC, Iwashyna TJ. Improving sepsis treatment by embracing diagnostic uncertainty. *Ann Am Thorac Soc*. 2019;16:426-429. <https://doi.org/10.1513/ANNALSATS.201809-646PS>
 21. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015;7:299ra122. https://doi.org/10.1126/SCITRANSLMED.AAB3719/SUPPL_FILE/7-299RA122_SM.PDF
 22. Wong A, Otlés E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181:1065-1070. <https://doi.org/10.1001/JAMAINTERNMED.2021.2626>
 23. Delahanty RJ, Alvarez JA, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Ann Emerg Med*. 2019;73:334-344. <https://doi.org/10.1016/J.ANNEMERGEMED.2018.11.036>
 24. Williams JM, Greenslade JH, McKenzie JV, Chu K, Brown AFT, Lipman J. Systemic inflammatory response syndrome, quick sequential organ function assessment, and organ dysfunction: insights from a prospective database of ED patients with infection. *Chest*. 2017;151:586-596. <https://doi.org/10.1016/j.chest.2016.10.057>
 25. Kijpaisalratana N, Sanglertsinlapachai D, Techaratsami S, Musikataavorn K, Saoraya J. Machine learning algorithms for early sepsis detection in the emergency department: a retrospective study. *Int J Med Inform*. 2022;160:104689. <https://doi.org/10.1016/j.ijmedinf.2022.104689>
 26. Johnson AEW, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc*. 2018;25:32-39. <https://doi.org/10.1093/JAMIA/OCX084>
 27. Rhee C, Dantes RB, Epstein L, Klompas M. Using objective clinical data to track progress on preventing and treating sepsis: CDC's new 'Adult Sepsis Event' surveillance strategy. *BMJ Qual Saf*. 2019;28:305-309. <https://doi.org/10.1136/BMJQS-2018-008331>
 28. Rhee C, Zhang Z, Kadri SS, et al.; CDC Prevention Epicenters Program. Sepsis surveillance using adult sepsis events simplified eSOFA criteria versus sepsis-3 SOFA criteria. *Crit Care Med*. 2019;47:307-314. <https://doi.org/10.1097/CCM.0000000000003521>
 29. Kalantari A, Mallema H, Weingart SD. Sepsis definitions: the search for gold and what CMS got wrong. *West J Emerg Med*. 2017;18:951-956. <https://doi.org/10.5811/WESTJEM.2017.4.32795>
 30. Venkatesh AK, Slesinger T, Whittle J, et al. Preliminary performance on the new CMS sepsis-1 national quality measure: early insights from the emergency quality network (E-QUAL). *Ann Emerg Med*. 2018;71:10-15.e1. <https://doi.org/10.1016/J.ANNEMERGEMED.2017.06.032>
 31. Johnson AEW, Aboab J, Raffa JD, et al. A comparative analysis of sepsis identification methods in an electronic database. *Crit Care Med*. 2018;46:494-499. <https://doi.org/10.1097/CCM.0000000000002965>
 32. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc*. 2020;27:1921-1934. <https://doi.org/10.1093/JAMIA/OCAA139>
 33. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy: a machine learning experiment to estimate treatment effects from randomized trial data. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005010. <https://doi.org/10.1161/CIRCOUTCOMES.118.005010>
 34. Marafino BJ, Schuler A, Liu VX, Escobar GJ, Baiocchi M. Predicting preventable hospital readmissions with causal machine learning. *Health Serv Res*. 2020;55:993-1002.
 35. Inoue K, Athey S, Tsugawa Y. Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. *Int J Epidemiol*. 2023;52:1243-1256. <https://doi.org/10.1093/IJE/DYAD037>
 36. Kamran F, Tang S, Otlés E, et al. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *BMJ*. 2022;376:e068576. <https://doi.org/10.1136/BMJ-2021-068576>
 37. Zabihi M, Kiranyaz S, Gabbouj M. Sepsis prediction in intensive care unit using ensemble of XGboost models. In: *2019 Computing in Cardiology Conference (CinC)*. 2019;45.
 38. Yang M, Wang X, Li Y. Early prediction of sepsis using multi-feature fusion based XGBoost learning and Bayesian optimization. <https://doi.org/10.22489/CinC.2019.020>

39. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci U S A*. 2019;116:4156-4165.
40. Xu Y, Ignatiadis N, Sverdrup E, Fleming S, Wager S, Shah NH. Treatment heterogeneity for survival outcomes. In: *Handbook of Matching and Weighting Adjustments for Causal Inference*. 2022:445-482. Published online. <https://doi.org/10.1201/9781003102670-21>
41. Xu Y, Bechler K, Callahan A, Shah N. Principled estimation and evaluation of treatment effect heterogeneity: a case study application to dabigatran for patients with atrial fibrillation. *J Biomed Inform*. 2023;143:104420. <https://doi.org/10.1016/j.jbi.2023.104420>
42. Saeed A, Mehta LS. Statin therapy in older adults for primary prevention of atherosclerotic cardiovascular disease: the balancing act—American College of Cardiology. Accessed December 19, 2023. <https://www.acc.org/Latest-in-Cardiology/Articles/2020/10/01/11/39/Statin-Therapy-in-Older-Adults-for-Primary-Prevention-of-Atherosclerotic-CV-Disease>
43. Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput Methods Programs Biomed*. 2020;196:105608. <https://doi.org/10.1016/j.cmpb.2020.105608>
44. Richardson S, Hirsch JS, Narasimhan M, et al.; the Northwell COVID-19 Research Consortium. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA*. 2020;323:2052-2059. <https://doi.org/10.1001/JAMA.2020.6775>
45. Booth A, Reed AB, Ponzo S, et al. Population risk factors for severe disease and mortality in COVID-19: a global systematic review and meta-analysis. *PLoS One*. 2021;16:e0247461. <https://doi.org/10.1371/JOURNAL.PONE.0247461>
46. Flaks-Manov N, Srulovici E, Yahalom R, Perry-Mezre H, Balicer R, Shadmi E. Preventing hospital readmissions: healthcare providers' perspectives on "impactibility" beyond EHR 30-day readmission risk prediction. *J Gen Intern Med*. 2020;35:1484-1489.
47. Mahase E. Covid-19: Pfizer's Paxlovid is 89% effective in patients at risk of serious illness, company reports. *BMJ*. 2021;375:n2713. <https://doi.org/10.1136/BMJ.N2713>
48. Kamran F. *Aligning Machine Learning Solutions with Clinical Needs*. Doctoral dissertation. University of Michigan; 2023.