Summer 8-15-2022

Development of the Assessment of Clinical Prediction Model Transportability (APT) Checklist

Sean Chonghwan Yu
Washington University in St. Louis

Recommended Citation

Yu, Sean Chonghwan, "Development of the Assessment of Clinical Prediction Model Transportability (APT) Checklist" (2022). *McKelvey School of Engineering Theses & Dissertations*. 811.
https://openscholarship.wustl.edu/eng_etds/811

WASHINGTON UNIVERSITY IN ST. LOUIS
McKelvey School of Engineering
Department of Biomedical Engineering

Dissertation Examination Committee:
Albert M. Lai, Chair
Dennis L. Barbour
Randi E. Foraker
Thomas G. Kannampallil
Philip R. O. Payne

Development of the Assessment of Clinical Prediction Model Transportability (APT) Checklist

by
Sean C. Yu

A dissertation presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2022
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
APACHE II	Acute Physiology And Chronic Health Evaluation II
ASE	Adult Sepsis Event
AUPRC	Area Under Precision Recall Curve
AUROC	Area Under Receiver Operating Characteristic curve
BJC	Barnes-Jewish Christian
BJH	Barnes-Jewish Hospital
CDC	Centers for Disease Control and Prevention
CDM	Common Data Model
CDS	Clinical Decision Support
CHARMS	Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
CMS	Centers for Medicare and Medicaid Services
COVID-19	COronaVIrus Disease 2019
CPM	Clinical Prediction Models
DES	Discrete-Event Simulation
ECMO	ExtraCorporeal Membrane Oxygenation
ED	Emergency Department
EHR	Electronic Health Records
eMERGE	Electronic Medical Records & Genomics
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resource
FiO2	Fraction of inspired Oxygen
GDBT	Gradient Boosted Decision Tree
GRASP	GRade and ASsess Predictive tools
HCP	HealthCare Process
HFOT	High Flow Oxygen Therapy
HITECH	Health Information Technology for Economic and Clinical Health
ICD	International Classification of Diseases
ICU	Intensive Care Unit

IMV	Invasive Mechanical Ventilation
IQR	InterQuartile Range
LFOT	Low Flow Oxygen Therapy
LogReg	Logistic Regression
LOINC	Logical Observation Identifiers Names and Codes
LOS	Length Of Stay
LR	Logistic Regression
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
NEWS	National Early Warning Score
NIMV	Non-Invasive Mechanical Ventilation
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PCR	Polymerase Chain Reaction
PGRN	Pharmacogenomics Research Network
POA	Present On Admission
PP	PathoPhysiological
PROBAST	Prediction model Risk Of Bias ASsessment Tool
QAD	Qualifying Antibiotic Days
qSOFA	quick SOFA
SaMD	Software as Medical Device
SHAP	SHapley Additive exPlanations
SIRS	Systemic Inflammatory Response Syndrome
SMART	Substitutable Medical Apps, Reusable Technologies
SNOMED	Systematized Nomenclature of Medicine
SOFA	Sequential Organ Failure Assessment
SOI	Suspicion Of Infection
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UMLS	Unified Medical Language System
XGB	eXtreme Gradient Boosted Trees Model

List of Appendices

Acknowledgements

I would like to thank the following people, all of whom were integral to me completing my dissertation and PhD:

Albert and Philip for giving me a chance to work in their lab and providing me with the opportunity to meet incredible people in the field, as well as for all of their mentorship not just in research, but also in career and in life.

The rest of the committee members – Randi, Tom, and Dennis – for offering invaluable advice and gently and patiently guiding me towards the finish line.

All of my partners and collaborators in research, without whom I could not have published – Drew, Aditi, Kevin, Mac, Marin, Ryan, Nirmala, Courtney, Po-Yin, and Pat.

All of the administrative staff at I2 and BME for taking good care of me – Andrea, Cynthia, Kim, Katie, and Isabelle among many others.

Countless faculty beyond my committee who generously shared with me their experience and advice – Shamim, Fuhai, Adam, Tiffany, and more.

Other trainees who kept me company and kept me sane throughout, especially during the last treacherous stretch – Mac, Josh, Nigel, Sayantan, Erin, Kriti, among many others.

My family for supporting me and encouraging me at all times – Chansu, Yongmee, and Andrew.

Finally, special thanks to Drew for being a patient subject matter expert, a reliable colleague, an excellent mentor, but most importantly, an invaluable friend. Thank you.

Sean C. Yu

Washington University in St. Louis

August 2022

ABSTRACT OF THE DISSERTATION

Development of the Assessment of Clinical Prediction Model Transportability (APT) Checklist

by

Sean C. Yu

Doctor of Philosophy in Biomedical Engineering

Washington University in St. Louis, 2022

Professor Albert M. Lai, Chair

Clinical Prediction Models (CPM) have long been used for Clinical Decision Support (CDS) initially based on simple clinical scoring systems, and increasingly based on complex machine learning models relying on large-scale Electronic Health Record (EHR) data. External implementation – or the application of CPMs on sites where it was not originally developed – is valuable as it reduces the need for redundant *de novo* CPM development, enables CPM usage by low resource organizations, facilitates external validation studies, and encourages collaborative development of CPMs. Further, adoption of externally developed CPMs has been facilitated by ongoing interoperability efforts in standards, policy, and tools. However, naïve implementations of external CPMs are prone to failure due to the incompatibilities between the environments of the development and implementation sites. Although prior research has described methods for estimating the external validity of predictive models, quantifying dataset shift, updating models, as well as numerous CPM-specific frameworks for guiding the development, evaluation, reporting, and systematic reviews of CPMs, there are no frameworks for assessing the compatibility between a CPM and the target environment. This dissertation addresses this critical gap by proposing a novel CPM transportability checklist for guiding the adoption of externally developed CPMs.

To guide the development of the checklist, four extant CPM-relevant frameworks (TRIPOD, CHARMS, PROBAST, and GRASP) were reviewed and synthesized, thereby identifying the key domains of CPMs. Then, four individual studies were conducted, each identifying, assessing the impact of, and/or proposing solutions for the disparity between CPM and environment in those domains. The first two studies target disparities in features, with the first characterizing the non-generalizability impact of a particular class of commonly used, EHR-idiosyncratic features. The second study was conducted to identify and propose a solution for the semantic discrepancy in features across sites caused by the insufficient coverage of EHR data by standards. The third study focused on the prediction target of CPMs, identifying significant heterogeneity in disease understanding, phenotyping algorithms, and cohort characteristics of the same clinical condition. In the fourth study investigating CPM evaluation, the gap between typical CPM evaluation design and expected implemented behavior was identified, and a novel evaluative framework was proposed to bridge that gap. Finally, the APT checklist was developed using the synthesis of the aforementioned CPM frameworks as the foundation, enriched through the incorporation of innovations and findings from these four conducted studies. While rigorous meta-evaluation remains, the APT checklist shows promise as a tool for assessing CPM transportability thereby reducing the risk of failure of externally implemented CPMs.

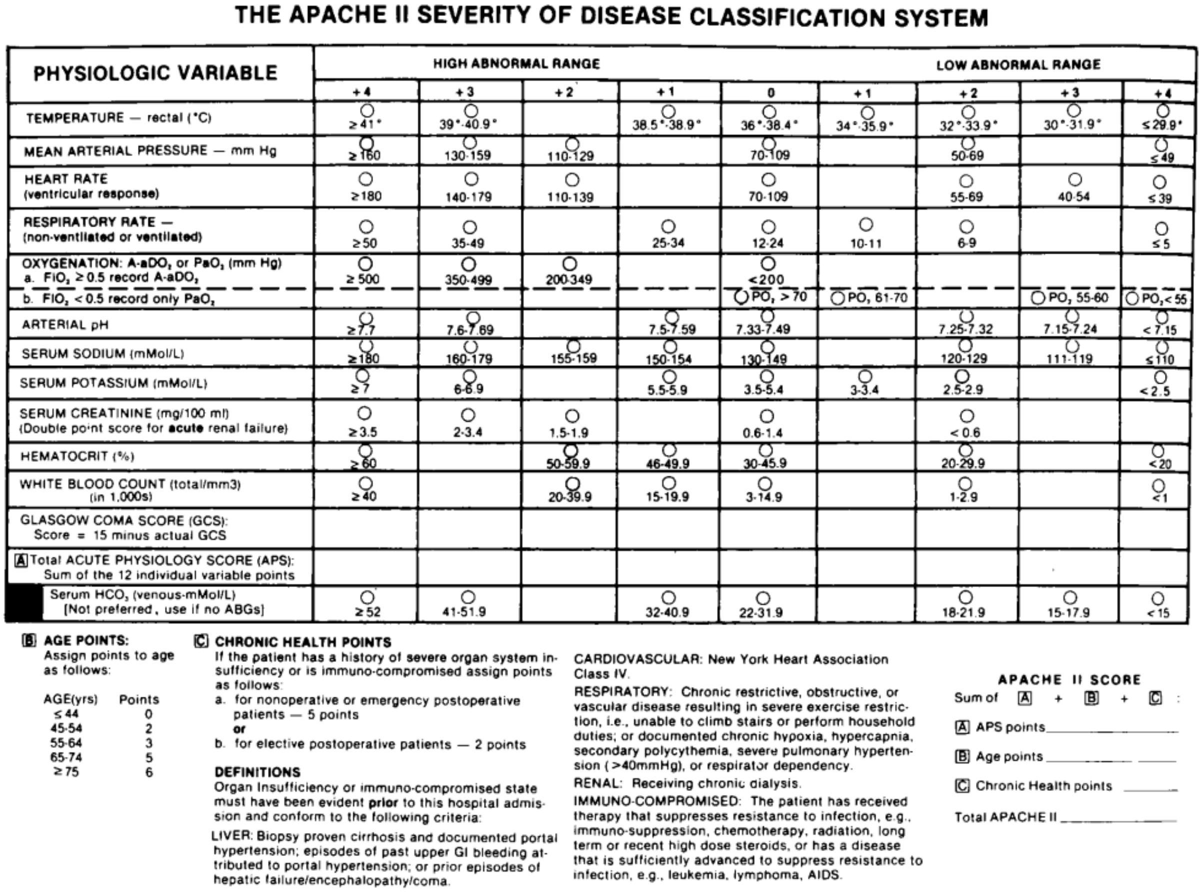
The key contributions to informatics include: the discovery of healthcare process (HCP) variables as a driver of CPM non-transportability, the fragility of clinical phenotyping used to identify CPM targets, a novel classification system and meta-heuristics for an aspect of EHR data previously lacking in standards, a novel CPM evaluation design termed the pseudo-prospective trial, and the APT checklist. Overall, this work contributes to the body of biomedical informatics literature guiding the success of informatics interventions.

Chapter 1. Introduction

1.1. General Background

1.1.1. Clinical Prediction Models

Clinical Prediction Models (**CPM**) use covariates to derive a risk score for individual patients that can be used to guide clinical decision-making. Based on the resultant score, healthcare providers may decide to pursue additional testing, or provide/withhold therapy.[1] This type of clinical decision support (**CDS**) has long been used in clinical practice – for instance, a commonly used illness severity score for predicting clinical deterioration in the intensive care unit (**ICU**) setting – Acute Physiology And Chronic Health Evaluation II (**APACHE II**) – was developed in 1985 (**Figure 1**).[2]



THE APACHE II SEVERITY OF DISEASE CLASSIFICATION SYSTEM

PHYSIOLOGIC VARIABLE	HIGH ABNORMAL RANGE					LOW ABNORMAL RANGE			
	+4	+3	+2	+1	0	+1	+2	+3	+4
TEMPERATURE — rectal (°C)	○ ≥41°	○ 39°-40.9°		○ 38.5°-38.9°	○ 36°-38.4°	○ 34°-35.9°	○ 32°-33.9°	○ 30°-31.9°	○ ≤29.9°
MEAN ARTERIAL PRESSURE — mm Hg	○ ≥160	○ 130-159	○ 110-129		○ 70-109		○ 50-69		○ ≤49
HEART RATE (ventricular response)	○ ≥180	○ 140-179	○ 110-139		○ 70-109		○ 55-69	○ 40-54	○ ≤39
RESPIRATORY RATE — (non-ventilated or ventilated)	○ ≥50	○ 35-49		○ 25-34	○ 12-24	○ 10-11	○ 6-9		○ ≤5
OXYGENATION: A-aDO$_2$ or PaO$_2$ (mm Hg) a. FiO$_2$ ≥ 0.5 record A-aDO$_2$	○ ≥500	○ 350-499	○ 200-349		○ <200				
b. FiO$_2$ < 0.5 record only PaO$_2$					○ PO$_2$ >70	○ PO$_2$ 61-70		○ PO$_2$ 55-60	○ PO$_2$ < 55
ARTERIAL pH	○ ≥7.7	○ 7.6-7.69		○ 7.5-7.59	○ 7.33-7.49		○ 7.25-7.32	○ 7.15-7.24	○ < 7.15
SERUM SODIUM (mMol/L)	○ ≥180	○ 160-179	○ 155-159	○ 150-154	○ 130-149		○ 120-129	○ 111-119	○ ≤110
SERUM POTASSIUM (mMol/L)	○ ≥7	○ 6-6.9		○ 5.5-5.9	○ 3.5-5.4	○ 3-3.4	○ 2.5-2.9		○ <2.5
SERUM CREATININE (mg/100 ml) (Double point score for **acute** renal failure)	○ ≥3.5	○ 2-3.4	○ 1.5-1.9		○ 0.6-1.4		○ < 0.6		
HEMATOCRIT (%)	○ ≥60		○ 50-59.9	○ 46-49.9	○ 30-45.9		○ 20-29.9		○ <20
WHITE BLOOD COUNT (total/mm3) (in 1,000s)	○ ≥40		○ 20-39.9	○ 15-19.9	○ 3-14.9		○ 1-2.9		○ <1
GLASGOW COMA SCORE (GCS): Score = 15 minus actual GCS									
A Total ACUTE PHYSIOLOGY SCORE (APS): Sum of the 12 individual variable points									
Serum HCO$_3$ (venous-mMol/L) [Not preferred, use if no ABGs]	○ ≥52	○ 41-51.9		○ 32-40.9	○ 22-31.9		○ 18-21.9	○ 15-17.9	○ <15

B AGE POINTS:
Assign points to age as follows:

AGE(yrs)	Points
≤44	0
45-54	2
55-64	3
65-74	5
≥75	6

C CHRONIC HEALTH POINTS
If the patient has a history of severe organ system insufficiency or is immuno-compromised assign points as follows:

a. for nonoperative or emergency postoperative patients — 5 points
or
b. for elective postoperative patients — 2 points

DEFINITIONS
Organ Insufficiency or immuno-compromised state must have been evident **prior** to this hospital admission and conform to the following criteria:

LIVER: Biopsy proven cirrhosis and documented portal hypertension; episodes of past upper GI bleeding attributed to portal hypertension; or prior episodes of hepatic failure/encephalopathy/coma.

CARDIOVASCULAR: New York Heart Association Class IV.

RESPIRATORY: Chronic restrictive, obstructive, or vascular disease resulting in severe exercise restriction, i.e., unable to climb stairs or perform household duties; or documented chronic hypoxia, hypercapnia, secondary polycythemia, severe pulmonary hypertension (>40mmHg), or respirator dependency.

RENAL: Receiving chronic dialysis.

IMMUNO-COMPROMISED: The patient has received therapy that suppresses resistance to infection, e.g., immuno-suppression, chemotherapy, radiation, long term or recent high dose steroids, or has a disease that is sufficiently advanced to suppress resistance to infection, e.g., leukemia, lymphoma, AIDS.

APACHE II SCORE
Sum of A + B + C :

A APS points ________________

B Age points ____________ ______

C Chronic Health points ________

Total APACHE II ______________

Figure 1. APACHE II

The Acute Physiology And Chronic Health Evaluation (APACHE) II is a relatively simple, expert/heuristic-derived clinical scoring system intended for use in the ICU for assessment of disease severity and risk stratification, validated by its ability to predict in-hospital mortality.

Recently, there has been a massive proliferation of CPMs due to the opportune confluence of three factors: 1) policy resulting in vast availability of clinical data; 2) free and accessible software facilitating the development of CPMs; and 3) availability of cheap compute. First, the Health Information Technology for Economic and Clinical Health (**HITECH**) act of 2009 accelerated the adoption of electronic health record (**EHR**) software and as a result, massively expanded the availability of clinical data for research.[3] Around the same time in 2012, a deep convolutional neural network called AlexNet achieved dramatic improvement on the ImageNet image classification challenge, triggering widespread excitement and optimism regarding

Artificial Intelligence (**AI**) and Machine Learning (**ML**), resulting in the proliferation of free, easy-to-use, and publicly available software facilitating the development of CPMs.[4] Meanwhile, computing power continued to get cheaper and easier to access. Combined, the availability of dense clinical data, inexpensive computing power, and AI/ML software resulted in a dramatic surge of clinical prediction model development efforts (**Figure 2**).[5]

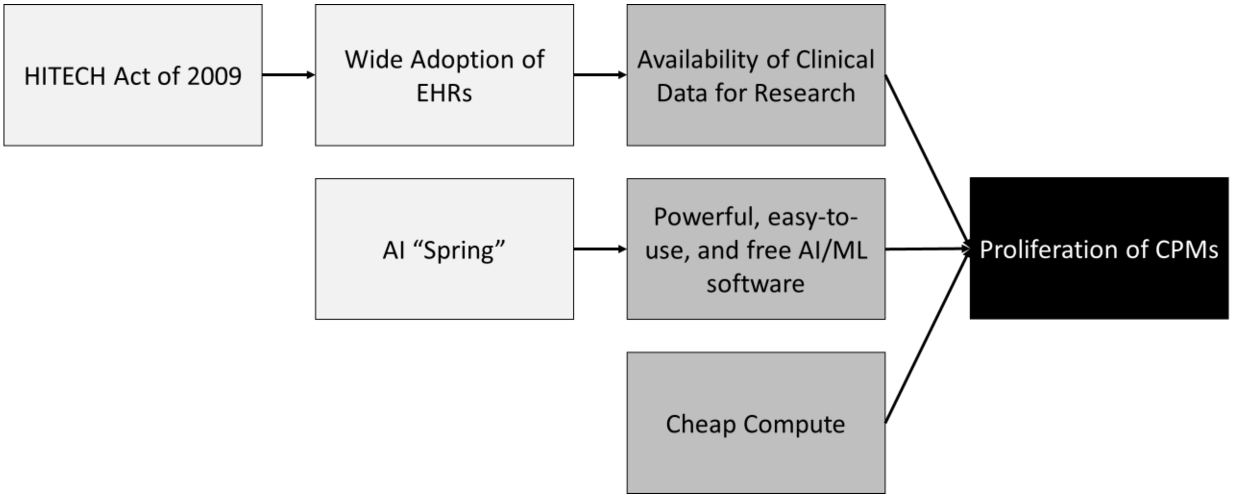

HITECH Act of 2009
Wide Adoption of EHRs
Availability of Clinical Data for Research
AI “Spring”
Powerful, easy-to-use, and free AI/ML software
Proliferation of CPMs
Cheap Compute

Figure 2. Causes of CPM Proliferation

The wide availability of clinical data, AI/ML software, and cheap compute has facilitated the development of CPMs, resulting in a massive proliferation of CPMs.

As a result, there are over 400 CPMs for chronic obstructive pulmonary disease outcomes, over 350 for cardiovascular disease risk, and over 200 for diagnosis and prognosis for a disease as recent as Coronavirus Disease 2019 (**COVID-19**) already by early 2020.[6-8] The overwhelming number of CPMs, many of which were designed for identical tasks, represents an enormously wasteful duplication of efforts.[5] Moreover, most of these models – many of which ostensibly outperform traditional scoring systems – are not evaluated externally, fewer are implemented into clinical practice, and fewer yet are adopted by organizations wherein the models were not originally developed.[9, 10]

1.1.2. External Implementation or Adoption of CPMs

The capacity to implement CPMs externally, or conversely, to implement externally developed CPMs (called adoption hereon), can directly contribute to the de-duplication of CPM development efforts by reducing the need for *de novo* development at each study site (**Figure 3**).[5] Further, external implementation or adoption of CPMs enable low-resource organizations to participate in the usage of CPMs, facilitate external validation studies, and overall, encourage the collaborative development of CPMs. In sum, external implementation or adoption has these numerous, impactful benefits but has been scarcely done for several reasons, a major one of which is the lack or insufficiency of health IT interoperability.[11]

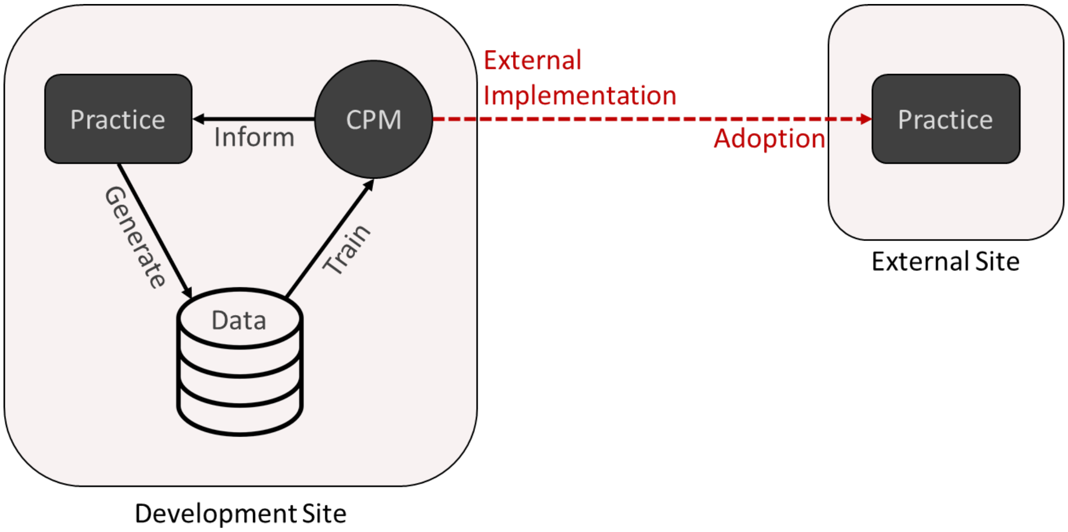


Figure 3. External Implementation or Adoption of CPMs

The Learning Health System (LHS) paradigm as introduced by the Institute of Medicine suggests that modern health systems should learn from data generated from their own practice to inform practice – i.e., train CPMs using EHR data to inform CDS. External implementation or adoption refers to the application of the CPM on an external site in which the CPM was not originally developed.

The inability or difficulty of differing health IT systems to exchange information has historically hampered external implementation or adoption of CPMs.[12] However, ongoing efforts in policy, standards, and tools increasingly facilitate external implementation or adoption of CPMs. Notably, the final rule for the 21st Century Cures Act require certified health IT to publish APIs using the HL7 Fast Healthcare Interoperability Resource (**FHIR**) standard. At the same time, Substitutable Medical Apps, Reusable Technologies (**SMART**) was developed as a FHIR-based application platform.[13] As a result, there are already site and EHR-agnostic CPM-based risk calculation applications hosted on SMART that can be adopted by any capable organization.[14] Maturation and adoption of interoperability standards and technology has and continues to improve the capacity of organizations to implement externally developed CPM, which increasingly presents as a competitive option to internal, *de novo* development.

Unfortunately, however, naïve external implementation or adoption of CPMs can result in significantly degraded performance.[15-17] Infamously, the Epic Sepsis Model was found to have an area under the receiver operating characteristic curve of 0.63 when tested externally compared to the interquartile range of 0.76 – 0.83 reported by the model developers.[18]

The causes of performance loss on external validation are extensive wide-ranging, some of which are well-known and well-studied – notably, the difference in population or case mix. For instance, the significant physiological difference between adult and pediatric populations results in the inefficacy of clinical scoring systems designed for adults in pediatric populations without critical modifications.[19, 20] Broadly, other considerations for CPM transportability include the technological capacity to faithfully reproduce the model, the human and financial resources to dedicate to the effort, the cultural willingness of the adopting organization to engage with the CPM, ethical and legal concerns surrounding the reliance of CPM-guided CDS, and how the technology will be integrated within the existing clinical workflow.[21-24] These general concerns overlap significantly with the discipline of technology acceptance and implementation science, in-depth discussions of which are out of scope for this dissertation.[25, 26] Specific to statistical modeling however, is the assessment of their ability to provide accurate results or predictions for an unseen population, or the study of model generalizability or transportability.[27]

1.1.3. Generalizability

A critical barrier to successful external implementation or adoption of CPMs is the problem of model non-generalizability, which has been studied not just in the context of CPMs but in the larger context of statistical learning or machine learning. Because prediction models are almost always intended to be deployed for samples they were not trained on – it's not useful to predict yesterday's weather – estimating or retaining high model performance on unforeseen samples is of critical importance and has received significant attention. What follows is a brief description of the types of inquiries in the domain of prediction model generalizability.

To structure the conversation on model generalizability, there have been developments in the conceptual frameworks of generalizability, including better defining and organizing concepts and terms regarding generalizability. One such type of study focuses on the types of generalizability or validity.[27] For example, Streyerberg first divides validity into internal and external, the latter of which is synonymous with generalizability or transportability, which can be further divided into temporal or geographic.[28] In the machine learning literature, the difference between training data and data seen during production has been termed data distribution drift or shift, which can be divided several categories including covariate shift, label shift, etc.[29] Further, there has been CPM specific literature investigating the categories and examples of dataset shift, as well as strategies for recognition and mitigation.[30]

There’s also a branch of study developing and testing methods for estimating the performance of the model on unseen, out-of-sample samples. The most well-known of these approaches is cross-validation, a re-sampling procedure akin to bootstrap in which a subset of the data is used to train the model and the remaining subset is used to test the model, and this process is repeated on different subsets of train and test data.[31] Methodological augmentations to cross-validation arose from investigations into temporality resulting in (rolling) temporal cross-validation, into data leakage resulting in patient as opposed to record-level splitting to avoid “identity confounding,” and into subgroup performance resulting in internal-external cross-validation (IECV).[32-34]

Monitoring and detecting non-generalizability post-deployment is especially important in the case of gradual temporal data drift. As such, there have been methods developed and compared on not just the identification of changes in the data, but also, if those identified changes have a meaningfully negative impact on performance.[35] In the field of biomedical informatics, specific to CPMs, there have been frameworks proposed to detect, for example, calibration drift over time.[36]

Once drift is detected and has been found to significantly degrade model performance, models can be updated to reduce performance loss. These updating methods fall under the discipline of domain adaptation or transfer learning and range from full re-training if there is sufficient target domain data and resources; or fine-tuning in the case of neural networks where only the parameters of the last few layers are trained; to simple linear recalibration.[37-40]

These model updating methods are *reactive* in that they require encountering data from the unforeseen domain and potentially experiencing model failure for some time before the need for updating is discovered. In response, recently, there has been a burgeoning field of study on *proactive* methods that aim to develop shift-stable or shift-resistant models.[41, 42]

1.1.4. Frameworks for CPMs

In parallel, there have been numerous frameworks proposed for guiding the reporting, systematic review, and comparative evaluation of CPMs. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (**TRIPOD**) checklist was developed to identify the minimum set of information necessary for critical appraisal of CPMs to guide the reporting of CPM studies, and consists of 22 items including eligibility criteria of participants, outcome definition, missing data handling methods, and model performance (**Appendix 1**).[43] While TRIPOD was designed for the reporting of individual studies, CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (**CHARMS**) and Prediction model Risk Of Bias ASessment Tool (**PROBAST**) were designed to guide systematic reviews of CPMs.[44, 45] CHARMS identifies 36 items to extract from individual studies to facilitate systematic reviews (**Appendix 2**). Unlike TRIPOD or CHARMS which focus on identifying the set of information needed to characterize CPMs, PROBAST focuses on the risk of bias and applicability of CPMs to systematic review questions, thus is more acutely concerned about the mismatch between CPMs and the use-case (**Appendix 3**). On the other hand, Grade and Assess Predictive tools (**GRASP**) was explicitly developed to enable comparative analysis of CPMs to facilitate selection of CPMs for implementation (**Appendix 4**).[46] As a tool more interested in implementation than TRIPOD, CHARMS, or PROBAST; GRASP focuses on phase of evaluation, usability, and potential or realized impact on either clinical effectiveness, patient safety, or healthcare efficiency. A brief comparative summary of these frameworks for enabling comparative evaluation of CPMs can be found in **Table 1**.

Table 1. Summary of CPM Frameworks

Category	Description	Frameworks			
		TRIPOD	CHARMS	PROBAST	GRASP
Background	Study rationale, scope, purpose, use-case	3a, 15b, 20			7, 8, 11, 19
Population	Data source and study design (RCT, registry, etc.)	4a, 4b	1.1	1.1	12
	Study setting (IP [ED, ICU, etc.], OP, etc.) including study period	5a	2.3		9
	Inclusion/eligibility criteria	5b	2.1	1.2	
	Population characteristics (including comparison when appropriate)	13b, 13c	2.2, 5.1, 8.3		
Target	Outcome/target definition	6a	3.1, 3.2, 3.3, 3.5, 3.6	3.1, 3.2, 3.4, 3.6	10
Modeling	Predictor descriptions (type, what, when, etc.)	7a	4.1, 4.2, 4.3	2.1	12, 13
	Missing data analysis and handling	9	6.1, 6.2, 6.3	4.4	
	Predictor manipulation/feature engineering	10a	4.5	4.2	
	Model type	10b	6.4, 6.8	4.6	15
	Model training procedure including feature selection	10b	6.6, 6.7	4.5	
	Model updating/recalibration	10e, 17	7.5		
Evaluation	Model evaluation procedure including performance metrics and calibration	10d, 16	7.1, 7.2, 7.3, 7.4, 8.1, 8.2	4.7, 4.8	
	Interpretation of results	18, 19a, 19b	9.1, 9.2		
Validation	Extent of validation (internal, external)	3b			16, 29, 30
	Usability				31
	Impact (clinical effectiveness, patient safety, healthcare efficiency)				32, 33, 34, 35

Synthesis of the following CPM frameworks through the merging and recategorization of common items: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD, **Appendix 1**), CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS, **Appendix 2**), Prediction model Risk Of Bias ASessment Tool (PROBAST, **Appendix 3**), and Grade and Assess Predictive tools (GRASP, **Appendix 4**) through merging and recategorization of common items.[43-46]

1.2. Problem or Gap in Literature

Despite the significant utility of external implementation or adoption, the associated critical challenges of non-transportability, the ongoing research on prediction model generalizability, and the development of various CPM-related frameworks, there are no CPM-specific frameworks for assessing the transportability of CPMs. In fact, one of the peer reviewers for the GRASP article addresses this very point, stating that post-implementation success is "highly susceptible to local differences between health systems, how they collect and store data, how they follow implementation protocols".[46] Thus the objective of this dissertation was to develop and propose a novel framework for assessing the compatibility between the CPM and the external environment in which it is to be implemented, called the **A**ssessment of clinical **P**rediction model **T**ransportability (**APT**) checklist.

1.3. Scope

The investigations that are part of this dissertation focus primarily on the data-rich acute-care setting, and on supervised machine learning models for binary classification trained using EHR data and used to estimate the probability of diagnosis or prognosis. Also, while there are many aforementioned factors critical to the success of CDS based on adopted CPMs such as the technical capacity of the adopting organization to faithfully re-create the model, this dissertation will focus on the modeling-specific categories of concerns as identified through the synthesis of extant CPM-related frameworks (**Table 1**). Among those categories, disparities in purpose or setting as a source of CPM non-transportability is well-known and well-studies, and thus will not be covered by this dissertation. Similarly, there has been significant and recent work investigating data drift monitoring and model updating specific to CPMs in the field of biomedical informatics by Sharon E. Davis, which thus will be out of scope for this dissertation.[36, 38] Finally, since the dissertation focuses on assessment of transportability pre-implementation, included studies will not focus on the validation category. Instead, this dissertation will focus on items pertaining to predictors under the modeling category, target, and evaluation.

1.4. Specific Aims

Each specific aim focuses on the remaining categories of concern- hereby called feature, target, and evaluation – discovering CPM-specific disparities between CPM and external environment, their impact on CPM transportability, and proposing solutions for mitigating non-transportability. The discoveries and innovations resulting from the specific aims are to be integrated into the synthesis of CPM frameworks (**Table 1**) to produce the APT checklist.

1.4.1. Aim 1 (Chapter 2): Feature Disparity

The first aim of the dissertation is: **To identify heterogeneity in features commonly used by and idiosyncratic to CPMs trained on EHR data, assess the impact of said heterogeneity on CPM transportability, and propose solutions for avoiding CPM non-transportability caused by feature disparity.** To this end, two studies were conducted. The first study investigates the non-generalizability impact of HealthCare Process (**HCP**) features which are CPM features specific to EHR data that are highly site-specific due to their dependence on hospital protocols, documentation culture, choice of hardware/software, etc. as opposed to features more directly based on underlying patient PathoPhysiology (**PP**). Through this study, it’s found that HCP features improve internal and temporal generalizability of CPMs to the detriment of external generalizability, and thus should be used with caution if at all. The second study is rooted in the idea of insufficient coverage of standardization in EHR data as a driving force for discrepancies in meaning and context of CPM features. In particular, respiratory support methods is a region of EHR data lacking in standards yet often used to generate features for CPMs, resulting in semantic heterogeneity of features based on respiratory support information. A novel classification system and accompanying, site-agnostic heuristics are proposed to mitigate the semantic discrepancy between sites. Together, these studies highlight sources of feature discrepancy unique to EHR data as well as methods to identify and mitigate their impact on CPM transportability.

1.4.2. Aim 2 (Chapter 3): Target Disparity

The second aim of the dissertation is: **To identify the causes and characterize the impact of heterogeneity in labels required for CPM development and to propose solutions for challenges to transportability of CPMs.** The targets of prediction for machine learning models in the acute care setting are clinical phenotypes commonly derived through rule-based criteria using EHR data. Heterogeneity in the overall disease concept understanding as well in the specific details of the phenotyping criteria give rise to heterogeneity in labeling, which can result in disagreements between the CPM development site and the external implementing site on who has or doesn't have the target phenotype, thereby limiting transportability. To study the extent of the impact of target label heterogeneity, a study was conducted investigating the fragility of sepsis phenotyping. Numerous well-validated and widely-used sepsis definitions and criteria were compared, finding significant differences in disease concept, criteria specifics, and resultant cohort characteristics as well as clinical outcomes. These findings highlight the critical importance of identifying not only the presence of, but also the potential impact of target label disparity on CPM transportability.

1.4.3. Aim 3 (Chapter 4): Evaluation Disparity

The third aim of the dissertation is: **To characterize and provide solutions for heterogeneity in the framing of CPM evaluation approaches by bridging the gap between CPM evaluation design and expected implemented behavior of CPM-based CDS.** Differences in the framing of CPM evaluation approaches can result in significantly different performance metrics between the CPM development site and the external implementing site beyond what can be explained by differences in case mix. The overall design of the CPM evaluation experiments – e.g., per-patient, per-encounter, or hourly – can significantly change the numeric values of performance metrics. Given that evaluation is performed to gauge real-world performance, CPM evaluation design are predicated on the expected implemented behavior of e.g., CDS based on CPM. Variability in evaluation design choices caused by differences in expected implementation behavior or otherwise give rise to disagreements among organizations on how a CPM ought to be evaluated, or put differently, how a model was evaluated and how it should have been evaluated. To enable an evaluative design more akin to real-world implemented behavior of CDS alerts based on CPM, a novel evaluative design, termed "pseudo-prospective trial," is proposed which facilitates the incorporation of factors unique to such settings. These factors include regularity or frequency of model execution, alert snoozing, dynamic inclusion/exclusion criteria, alternate/surrogate/competing outcomes at various time horizons, and more. The pseudo-prospective trial concept was proposed, developed, and demonstrated in a study using sepsis prediction in the general ward setting as the clinical context. The novel pseudo-prospective trial framework was found to significantly expand and enhance understanding of CPM performance, shows promise to reduce disparity between CPM evaluation approach at the CPM development site and expected implemented behavior at the external implementing site, thereby reducing the risk of CPM non-transportability.

1.5. Overview of Dissertation Structure

To summarize, this dissertation describes the research effort regarding the development of a checklist intended for use by the implementing organization on the formal assessment of CPM transportability, or the compatibility between the CPM and the external environment in which it is to be implemented. This dissertation in organized into five chapters, the first of which is this introductory chapter. Each of the following three chapters address each of the aforementioned specific aims. The fifth, penultimate chapter synthesizes the findings of the preceding three chapters and presents the APT checklist. The sixth and final chapter contains the summary and conclusions. In more detail, the chapters are as follows:

1. This first introductory chapter lays out the background, gap in literature or motivating problem, scope, specific aims, and the overall structure of the dissertation.
2. The second chapter addresses specific aim 1 – feature disparity – investigating the barriers of CPM transportability regarding the inputs of CPMs, conducted through two studies. The first study focuses on identifying the generalizability impact of HCP features, finding them to improve internal performance at the cost of harming external generalizability. The second study proposes a solution for a major category of cause for feature disparity – lack of standards – in the domain of respiratory support methods in order to mitigate the risk of semantic discrepancy.
3. The third chapter addresses specific aim 2 – target disparity – investigating the barriers of CPM transportability regarding the target of CPMs. A study is conducted in the clinical domain of sepsis, assessing the heterogeneity of overarching disease concept, specifications of phenotyping criteria, and the resultant cohort characteristics including clinical outcomes. The finding highlights the fragility of clinical phenotyping approaches, and the potential impact of disagreement on target phenotyping approaches on CPM generalizability.

4. The fourth chapter addresses specific aim 3 – evaluation disparity – investigating the barriers of CPM transportability regarding the evaluative design of CPMs. A study is conducted in which a novel CPM evaluation framework called the pseudo-prospective trial is proposed, developed, and demonstrated using sepsis prediction in the general ward as the clinical context. The pseudo-prospective trial shows promise as a framework for facilitating parity between CPM evaluation design and expected implemented behavior.
5. The fifth and penultimate chapter synthesizes the findings of the preceding three chapters into the primary contribution of this dissertation, the APT checklist.
6. The sixth and final chapter is comprised of the summary and concluding remarks.

The structure of the dissertation is provided in a graphical form as follows:

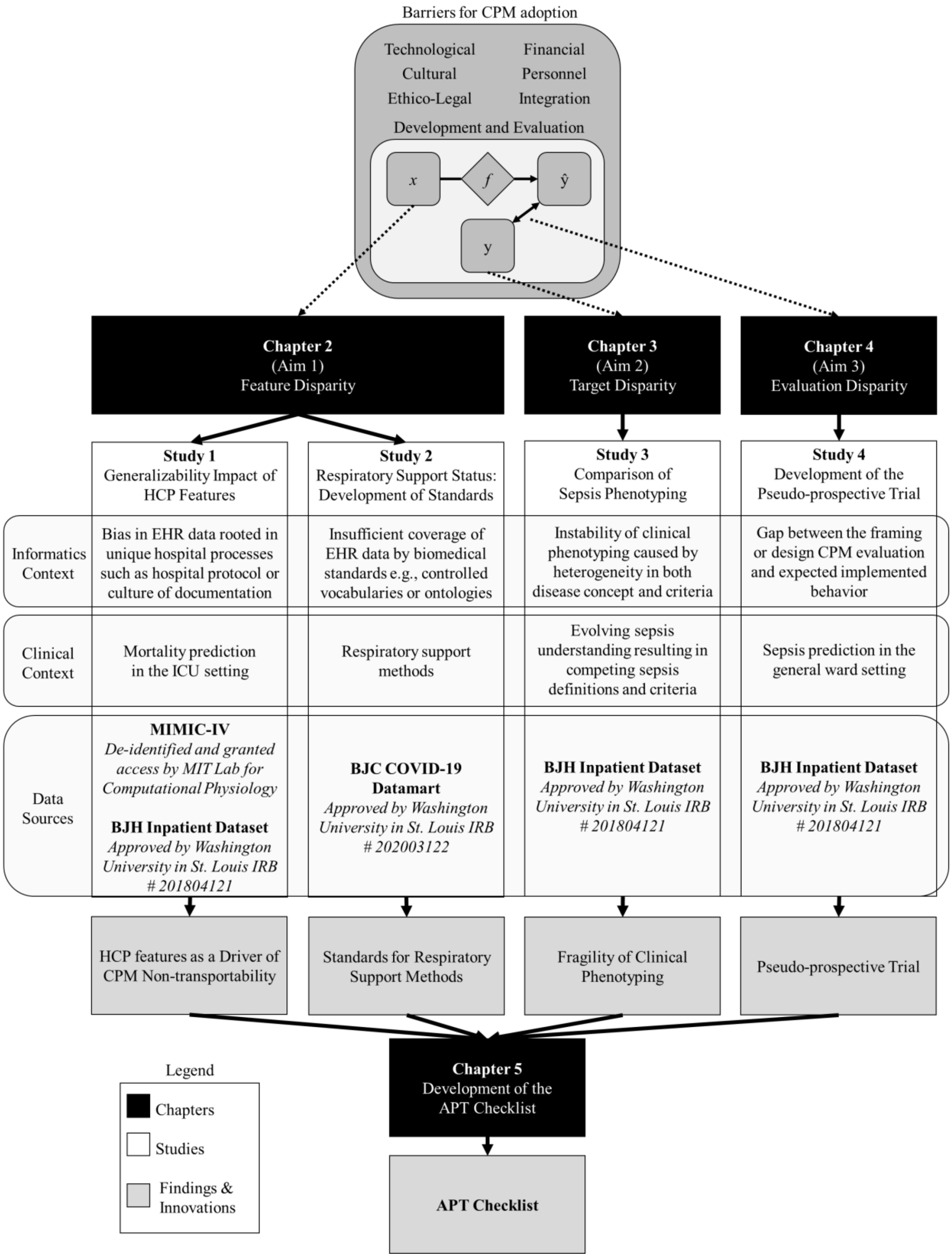

Barriers for CPM adoption
Technological
Financial
Cultural
Personnel
Ethico-Legal
Integration
Development and Evaluation
x
f
ŷ
y
Chapter 2
(Aim 1)
Feature Disparity
Chapter 3
(Aim 2)
Target Disparity
Chapter 4
(Aim 3)
Evaluation Disparity
Study 1
Generalizability Impact of HCP Features
Study 2
Respiratory Support Status: Development of Standards
Study 3
Comparison of Sepsis Phenotyping
Study 4
Development of the Pseudo-prospective Trial
Informatics Context
Bias in EHR data rooted in unique hospital processes such as hospital protocol or culture of documentation
Insufficient coverage of EHR data by biomedical standards e.g., controlled vocabularies or ontologies
Instability of clinical phenotyping caused by heterogeneity in both disease concept and criteria
Gap between the framing or design CPM evaluation and expected implemented behavior
Clinical Context
Mortality prediction in the ICU setting
Respiratory support methods
Evolving sepsis understanding resulting in competing sepsis definitions and criteria
Sepsis prediction in the general ward setting
Data Sources
MIMIC-IV
De-identified and granted access by MIT Lab for Computational Physiology
BJH Inpatient Dataset
Approved by Washington University in St. Louis IRB # 201804121
BJC COVID-19 Datamart
Approved by Washington University in St. Louis IRB # 202003122
BJH Inpatient Dataset
Approved by Washington University in St. Louis IRB # 201804121
BJH Inpatient Dataset
Approved by Washington University in St. Louis IRB # 201804121
HCP features as a Driver of CPM Non-transportability
Standards for Respiratory Support Methods
Fragility of Clinical Phenotyping
Pseudo-prospective Trial
Chapter 5
Development of the APT Checklist
APT Checklist
Legend
Chapters
Studies
Findings & Innovations

Figure 4. Dissertation Overview

Overview of the dissertation in graphical form. Chapters 2, 3, and 4 addresses specific aims 1, 2, and 3 respectively, and chapter 5 describes the development of the APT checklist, the primary contribution of this dissertation.

Chapter 2. Feature Disparity

2.1. Introduction

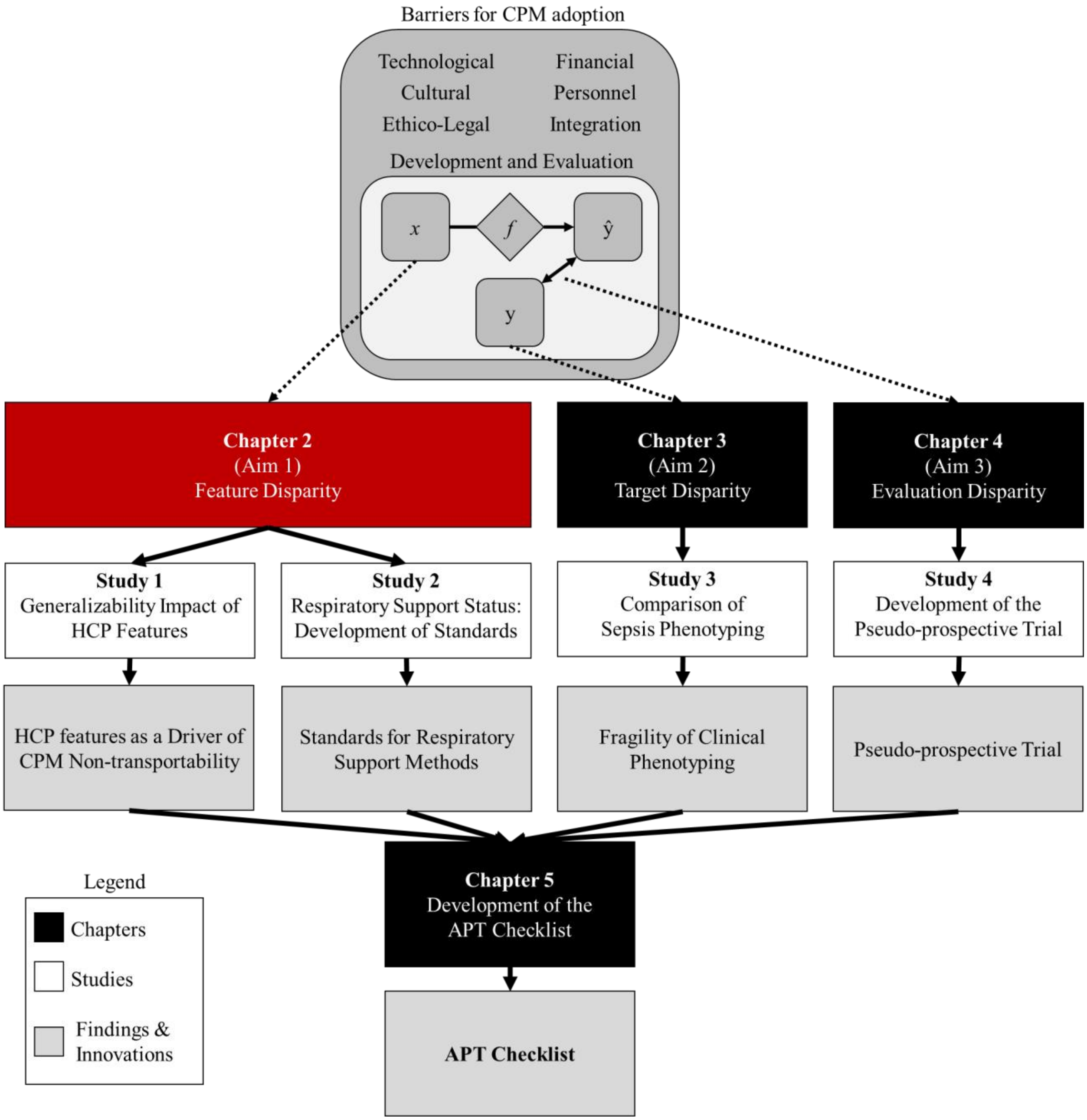

Barriers for CPM adoption
Technological
Financial
Cultural
Personnel
Ethico-Legal
Integration
Development and Evaluation
x
f
ŷ
y
Chapter 2
(Aim 1)
Feature Disparity
Chapter 3
(Aim 2)
Target Disparity
Chapter 4
(Aim 3)
Evaluation Disparity
Study 1
Generalizability Impact of HCP Features
Study 2
Respiratory Support Status: Development of Standards
Study 3
Comparison of Sepsis Phenotyping
Study 4
Development of the Pseudo-prospective Trial
HCP features as a Driver of CPM Non-transportability
Standards for Respiratory Support Methods
Fragility of Clinical Phenotyping
Pseudo-prospective Trial
Chapter 5
Development of the APT Checklist
APT Checklist
Legend
Chapters
Studies
Findings & Innovations

Figure 5. Chapter 2 Overview

This chapter addresses specific aim 1, the objective of which is to identify heterogeneity in features commonly used by and idiosyncratic to CPMs trained on EHR data, assess the impact of said heterogeneity on CPM transportability, and propose solutions for avoiding CPM non-transportability caused by feature disparity. To this end, two studies are conducted.

The first study assesses the generalizability impact of features unique to EHR data and commonly used in CPMs called HCP features which are heavily influenced by site-specific idiosyncrasies such as hospital protocols, documentation culture, choice of software/hardware, etc. It has been found that while HCP features improve estimated out-of-sample performance as measured through cross-validation, they harm external (cross-site) performance. Thus, those seeking to adopt a CPM and are concerned about CPM transportability should first assess if the CPM relies heavily on HCP features, and if so, develop a plan for handling HCP features or exclude the CPM from consideration. In addition, the insufficient coverage of EHR data by standards such as controlled vocabularies or ontologies can result in disagreements on the meaning or semantics of features and/or how to identify or derive those features. This feature ambiguity can severely limit the ability of the CPM adopting organization to faithfully replicate the model, thereby limiting transportability.

A second study was performed, identifying respiratory support methods as a domain of EHR data lacking in standards, proposing a novel standard, as well as developing an accompanying set of EHR-agnostic heuristics for identifying respiratory support episodes from raw EHR data. The innovations of the study contribute to the work of standardization or harmonization of EHR data, and shows promise as tools for reducing the semantic heterogeneity of EHR data. Because semantic heterogeneity limits CPM transportability and adherence to standards reduce semantic heterogeneity, CPM adopters should prefer adherence features to standards when possible, or the documentation of an ad-hoc standard when not.

The findings and innovations of the two studies that comprise this chapter are then used to supplement the development of the APT checklist.

2.2. Overview

Structurally, this chapter will begin by reiterating the motivating specific aims as has already been done in the preceding section, followed by a background section on the following topics – data drift/shift and data harmonization. Then, the two studies that comprise this chapter are presented, each including their own introduction, background, methods, results, discussion, and conclusion sections. Finally, the chapter concludes by discussing the ramifications of the findings and innovations of the studies on CPM transportability and the APT checklist.

2.3. Background and Significance

2.3.1. Data Shift

When the relationship between model inputs – also known as predictors or covariates – and outputs – also known as target – change between two datasets, e.g., training vs. testing, the machine learning community has converged on calling this phenomenon dataset drift or dataset shift.[29] When this dataset shift occurs between the CPM development site and an external implementing site, it can result in significantly degraded performance in the adopting site, thus is a barrier to CPM transportability and of interest to the APT checklist.[47]

There are numerous potential types and causes of dataset shift such as prior probability shift in which the distribution of the outcome or target variable differ between environments.[48] A comprehensive treatment of all possible reasons behind dataset shift in machine learning is out of scope for this dissertation, however, interested readers are encouraged to read *Dataset Shift in Machine Learning* by Quiñonero-Candela et al.[29]

Interestingly, the clinical discipline frames the problem of dataset shift slightly differently, borrowing terms and concepts from the study of clinical interventions.[49] Clinical trialists are concerned with the generalizability of clinical trial results – that the findings of Randomized Controlled Trials (**RCT**) are in fact applicable to patients beyond or external to the RCT population or not due to risks of bias.[50, 51] As CPM research lies in the intersection of clinical research and ML research, there have been recent efforts in unifying the "two worlds" into a unified framework of generalizability.[49]

There are many potential causes of data shift in clinical prediction modeling, many of which are not specific to CPMs and are relevant to ML research in general. However, the focus of this dissertation and chapter is on the causes of data shift unique to CPM using EHR data, beyond those caused by differences in patient population characteristics. The first study investigates a particular cause of data shift that is idiosyncratic to EHR data, and the impact of that shift on CPM generalizability. The second study focuses on a data shift stemming from the inadequacy of clinical or EHR data harmonization efforts.

2.3.2. Clinical Data Interoperability and Harmonization

When different healthcare organizations differ in the structure or syntax of clinical data as well as the meaning or semantics of data elements that are ostensibly identical, it limits the ability for a CPM adopting organization to faithfully recreate features as it was done by the development site, thereby causing dataset shift and acting as a barrier to CPM transportability. The process of unifying the representation of data from multiple sources so as to minimize problems caused by the discrepancy in data representation among sites is called data normalization or data harmonization.[52] Broadly, there are two dimensions of harmonization – syntactic and semantic.[53]

Syntactic harmonization and interoperability is enabled by the proliferation and adoption of data exchange standards as those described in chapter 1, such as the health messaging standards developed by Health Level 7 (**HL7**) including FHIR.[54] The usage of such data transmission standards which provide the structure of information packages enable different organizations to exchange clinical data without encountering syntax errors.[11, 55]

Syntactic interoperability, however, does not guarantee semantic interoperability – the ability for different organizations to understand the context and meaning of exchanged information.[11] Harmonization on this sematic level is accomplished using standards such as controlled vocabularies, terminologies, or ontologies.[11, 56] Various aspects of clinical EHR data have their own corresponding standards – for example, diagnosis information in the United States across health systems is predominantly coded using International Classification of Diseases (**ICD**) codes.[57] Further, there are data models that enforce mapping of various sections of EHR data to standards – for example, under the Observational Health Data Sciences and Informatics (**OHDSI**) Observational Medical Outcomes Partnership (**OMOP**) Common Data Model (**CDM**), conditions and procedures are mapped to ICD, measurements are mapped to Logical Observation Identifiers Names and Codes (**LOINC**), drugs are mapped to RxNorm, and so on.[58, 59]

While the use of these standards have enabled semantic interoperability where applicable, there remain vast domains of EHR data where standards are not available, not widely used, applied improperly, or where there are numerous competing standards.[60, 61] For example, LOINC is commonly used as the standard for all “measurements” such as lab results, vital signs, cultures, etc. but its use is highly inconsistent so as to limit interoperability.[62]

Despite ongoing herculean efforts in the development and dissemination of standards for EHR data, significant portions of EHR data remain without standards. These regions of EHR data lacking in standards are, nonetheless, used to generate features due to their information content that can be exploited to improve CPM performance. However, CPMs using these non-standard features are difficult to recapitulate thus limiting CPM transportability. The second study of this chapter focuses on respiratory support as an exemplary case where a domain of EHR data is commonly used for CPMs but is lacking in standards.

2.4. Study 1: Generalizability Impact of Healthcare Process (HCP) Features

This study assesses the generalizability impact of features unique to EHR data and commonly used in CPMs called HCP features which, unlike pathophysiological features, are heavily influenced by site-specific idiosyncrasies such as hospital protocols, documentation culture, choice of software/hardware.

2.4.1. Introduction

The widespread adoption of electronic health records (EHR), lowering cost of computing power, and accessible statistical software have enabled the proliferation of CPMs. Typically, CPMs are trained using various aspects of EHR data such as demographics, vital signs, and lab results to predict some clinical outcome such as disease onset, deterioration, or death.

EHR data is observational in that it is generated as part of routine clinical care, not directly for research, thus is subject to quality issues and biases. The source of these problems may be external – for example, diagnosis documentation is influenced by billing or financial reimbursement.[63-65] The other core cause is the variable process of documentation. For example, clinical lab tests are only performed when appropriate for the patient's pathophysiology and also, only if the clinician decides to order them. Clinician behavior on when and how to reveal and record the underlying patient state is influenced by policy, protocol, and culture – what some have called the healthcare process model or clinician-initiated data.[66, 67] In other words, EHR data is not a direct representation of the true underlying patient state, but rather, filtered through the lens of the healthcare process, thus is influenced by the idiosyncrasies of healthcare processes.

Of the manifestations of HCP variability, one major concern for CPM developers is nonrandom and heterogeneous data missingness. A common strategy to take advantage of informative missingness is to add indicator variables as features.[68] In fact, one study used only timestamps (rather than the actual measurement values) to develop a CPM, and found the model to have an AUROC of 0.707 in predicting clinical deterioration.[69] Beyond missingness, HCP can impact timing, frequency, or rate of measurements, each of which could be captured as features to further augment prediction.[70, 71] One study found that for clinical lab tests, HCP variables are often more informative than the actual measurement values, which they called pathophysiological variables.[71] In summary, manifestations of HCP such as missingness, timing, and frequency of measurements are clinically meaningful and informative, thus the incorporation of HCP variables into CPM is likely to improve performance.

However, given the heterogeneity of healthcare processes across institutions, the information content captured in HCP variables may also vary, resulting in low generalizability of CPMs relying on HCP as opposed to PP features. For example, in a comparison of publicly available ICU datasets, one study found that the mean number of heart rate measurements per hour varied drastically from roughly 1 to 30.[72] Thus, given the diversity of HCP including physician behavior profiles, it has been hypothesized that CPMs heavily reliant on these signals will have limited generalizability.[67] Futoma et al. have explicitly tested this hypothesis through assessing the impact of missingness indicator variables on CPM generalizability – they found that CPMs solely using indicator variables and those using indicator and physiologic variables, both failed to generalize across different study sites.[73] However, their study was limited to missingness indicator variables, did not interrogate temporal generalizability, and only used logistic regression. The objective of this study was to expand the understanding of HCP variables on generalizability through extending HCP variable types, modes of generalizability, and model types.

2.4.2. Methods

2.4.2.1. Study Design, Data Sources, and Population

Two data sources were used for this study; 1) Medical Information Mart for Intensive Care (**MIMIC)** IV, a publicly available, de-identified critical care database based on a tertiary academic medical center in Boston, MA, USA; and 2) BJH, an EHR dataset consisting of patients admitted to a tertiary academic medical center in St. Louis, MO, USA.[74] For both cohorts, only ICU stays were used for the analysis, which had to be at least 24 hours long and at most 30 days in duration. ICU stays with less than a 24-hour gap were merged as a single ICU stay. ICU stays were only included if patient was at least 18 years old at time of ICU stay start.

Patients missing critical EHR data such as admission-transfer-discharge, demographics, vital signs, and lab results data were excluded. To exclude patients with significant and unusual missingness especially in vital signs and lab results, only ICU stays with at least 12 heart rate measurements, 1 white blood cell count, and 1 creatinine measurement were included for analysis. Only the first ICU stay per patient were used for analysis to avoid “identity confounding”.[75] All dates within MIMIC-IV are shifted for anonymization purposes, though the shift is consistent for each patient. However, MIMIC-IV does effectively provide a range within which the real date may lie, so a random date within the possible time window was assigned. To accomplish parity in time periods between the two datasets, encounter admission dates were limited to between 1/1/12 and 12/31/19. Each dataset was split at roughly the midpoint, 6/1/15, into two eras: 0 (pre-split) and 1 (post-split); resulting in four final cohorts: MIMIC-IV era 0, MIMIC-IV era 1, BJH era 0, and BJH era 1. For all cohorts, the following were extracted from their respective EHR data: demographics, admission-transfer-discharge or patient location, vital signs, diagnoses, and lab results. The task for all models/experiments was to predict in-hospital mortality within 30 days of ICU admission using information from the first 24 hours of ICU admission.

2.4.2.2. Pathophysiological and Healthcare Process Feature-Sets

In accordance to the framework laid out by Hripcsak and Albers[66], features for prediction were divided into two types: pathophysiological and healthcare process (**Table 2**). Pathophysiological features included demographics, vital sign values, and lab result values (and transformations thereof such as median, interquartile range, etc.). Healthcare process features included day-of-week and time-of-day of admission and number of measurements per lab or vital sign. Number of measurements were also stratified by six subdivisions of the day (akin to shifts). Further, as there were physiologically infeasible outliers (3 median absolute deviation greater than the 97.5th percentile or lesser than the 2.5th percentile) in the data, the number of outlier measurements per lab or vital sign were also included as features. Occasionally, lab results and vital signs entries or rows contained null result values, which is often the case when the measurement was faulty or incomplete, so the number of null measurements per lab or vital sign were also included as features. All compared models were trained using either only pathophysiological features (PP), only healthcare process features (HCP), or both (HCP + PP). Features where the vast majority (>95%) of responses were identical or missing were removed as they were less likely to be informative.

Table 2. Pathophysiological vs. Healthcare Process Feature-Sets

Feature-Set	Definition	Example Features
Pathophysiology (PP)	Direct measure of patients' true state	Age at admission, median bilirubin, median absolute deviation of heart rate, 75th percentile of SpO2
Healthcare Process (HCP)	Indirect measure of patients' true state, influenced by the recording process	Number of creatinine measurements, number of POC glucose measurements (3rd shift), admission day of week (Saturday), number of FiO2 null measurements, number of outlier respiratory rate measurements

[a]Time of day was divided into six 4-hour shifts

2.4.2.3. Compared Models

Three different models were compared: 1) vanilla logistic regression (**LR**); 2) logistic regression with univariate basis spline feature expansion and hyperparameter optimization (**SplineLR opt**); and 3) XGBoost with hyperparameter optimization (**XGB opt**). For all three models, the model pipeline included a standardization step (zero-mean, unit-variance), and an imputation step (median). For SplineLR opt and XGB opt, hyperparameters were optimized using random search

(100 iterations of 5-fold CV), optimizing for maximum negative log loss. The regularization strength parameter was optimized for SplineLR, whereas the learning rate in addition to regularization strength parameter was optimized for XGB opt (**Appendix 5, Appendix 6**).

2.4.2.4. Experimental Design and Evaluation

Each data source-era combination was subsampled to have the same outcome prevalence to enable direct comparison of log loss values. Then, for each data source, era, feature set, and model type, a model was trained. The internal performance of the model was evaluated through the test log loss of three replicates of 5-fold cross validation. Then the models were applied externally cross-site and cross-time. For example, a model trained on BJH era 0 was evaluated on BJH era 1 (cross-time), MIMIC era 0 (cross-site), and MIMIC era 1 (cross-site and time). External performance assessed using 15 bootstrap iterations to match the total number of iterations of the internal evaluation procedure, and also because excessive iterations can yield overly narrow distributions that are overconfident as they do not account for sampling variability.[76] First, the internal performance of different model types were compared across feature-sets and site-era combinations. Second, the internal performance of different feature-sets were compared across model types and site-era combinations. Third, the difference between internal and external performances were compared across model types, site-eras, and feature-sets (**Table 3**).

Table 3. Dimensions of Comparison

Dimension	**Members**
Temporal	• Pre-7/1/2015 (**Era 0**) • Post-7/1/2015 (**Era 1**)
Site	• Barnes-Jewish Hospital (**BJH**) • Medical Information Mart for Intensive Care (**MIMIC-IV**)
Model type	• Vanilla logistic regression (**LR**) • Logistic regression with basis spline feature expansion with hyperparameter optimization (**SplineLR opt**) • XGBoost with hyperparameter optimization (**XGB opt**)
Feature-set	• Pathophysiological (**PP**) • Healthcare process (**HCP**)

In this study, the impact of feature-set on temporal and external (site) generalizability of clinical prediction models was evaluated, and how model type affected loss of generalizability.

2.4.3. Results

2.4.3.1. Cohort Characterization

The cohorts varied in size from 18,482 of MIMIC-IV 1 to 26,110 of BJH 1 (**Table 4**). The MIMIC-IV cohorts tended to be slightly older than the BJH cohorts (median age of 66 and 67 vs. 61 and 62). All cohorts had similar proportions of males ranging from 56.0% to 57.2%. While all cohorts had a similar proportion of white patients ranging from 65.7% to 69.9%, BJH had a higher proportion of black patients (23.9% and 25.6% of BJH vs. 10.8% and 10.4% of MIMIC-IV), and notably, BJH had no Hispanic population because Hispanic was not a valid response for race in BJH. BMI was similar across cohorts ranging from 27.0 to 27.8, as was ICU length of stay (**LOS**) from 2.4 to 2.8, whereas total LOS varied from 7.6 of MIMIC-IV 0 to 9.3 of BJH 0. The 30-day in-hospital mortality rate ranged from 8.2% of BJH 0 to 10.5% of BJH 1.

Table 4. Cohort Comparison

Variable	BJH 0 (n = 19,910)	BJH 1 (n = 26,110)	MIMIC-IV 0 (n = 17,273)	MIMIC-IV 1 (n = 18,482)
Age, median (IQR)	61 (51 – 72)	62 (50 – 71)	66 (55 – 78)	67 (56 – 77)
Sex (male), n (%)	11,209 (56.3%)	14,710 (56.3%)	9,676 (56.0%)	10,569 (57.2%)
Race, n (%)				
White, n (%)	13,916 (69.9%)	17,935 (68.7%)	11,615 (67.2%)	12,143 (65.7%)
Black, n (%)	4,753 (23.9%)	6,692 (25.6%)	1,857 (10.8%)	1,929 (10.4%)
Hispanic, n (%)	0 (0.0%)	0 (0.0%)	704 (4.1%)	687 (3.7%)
Asian, n (%)	117 (0.6%)	189 (0.7%)	473 (2.7%)	554 (3.0%)
Other/unknown, n (%)	1,124 (5.6%)	1,294 (5.0%)	2,624 (15.2%)	3,169 (17.1%)
BMI, median (IQR)	27.8 (23.7 – 33.2)	27.8 (23.6 – 33.5)	27.0 (23.9 – 33.0)	27.2 (23.8 – 32.0)
ICU LOS (days), median (IQR)	2.6 (1.7 – 5.0)	2.8 (1.7 – 5.1)	2.4 (1.5 – 4.3)	2.6 (1.6 – 4.8)
LOS (days), median (IQR)	9.3 (5.5 – 16.9)	9.0 (5.3 – 16.9)	7.6 (4.7 – 13.1)	8.1 (5.0 – 14.6)
30-day in-hospital mortality, n (%)	1,638 (8.2%)	2,753 (10.5%)	1,657 (9.6%)	1,873 (10.1%)

Cohort characteristics and clinical outcomes of four cohorts, the combination of two study sites and two time periods.

2.4.3.2. Comparison of Model Types for Internal Performance

The internal cross-validation (test) model performance of different model types (LR, SplineLR opt, and XGB opt) were compared, repeated for every site, era, and feature-set combination (**Figure 6**). In every combination, SplineLR opt and XGB opt outperformed LR (**Table 5**). However, in every combination, the difference in performance between SplineLR opt and XGB opt were not significant (**Figure 6**).

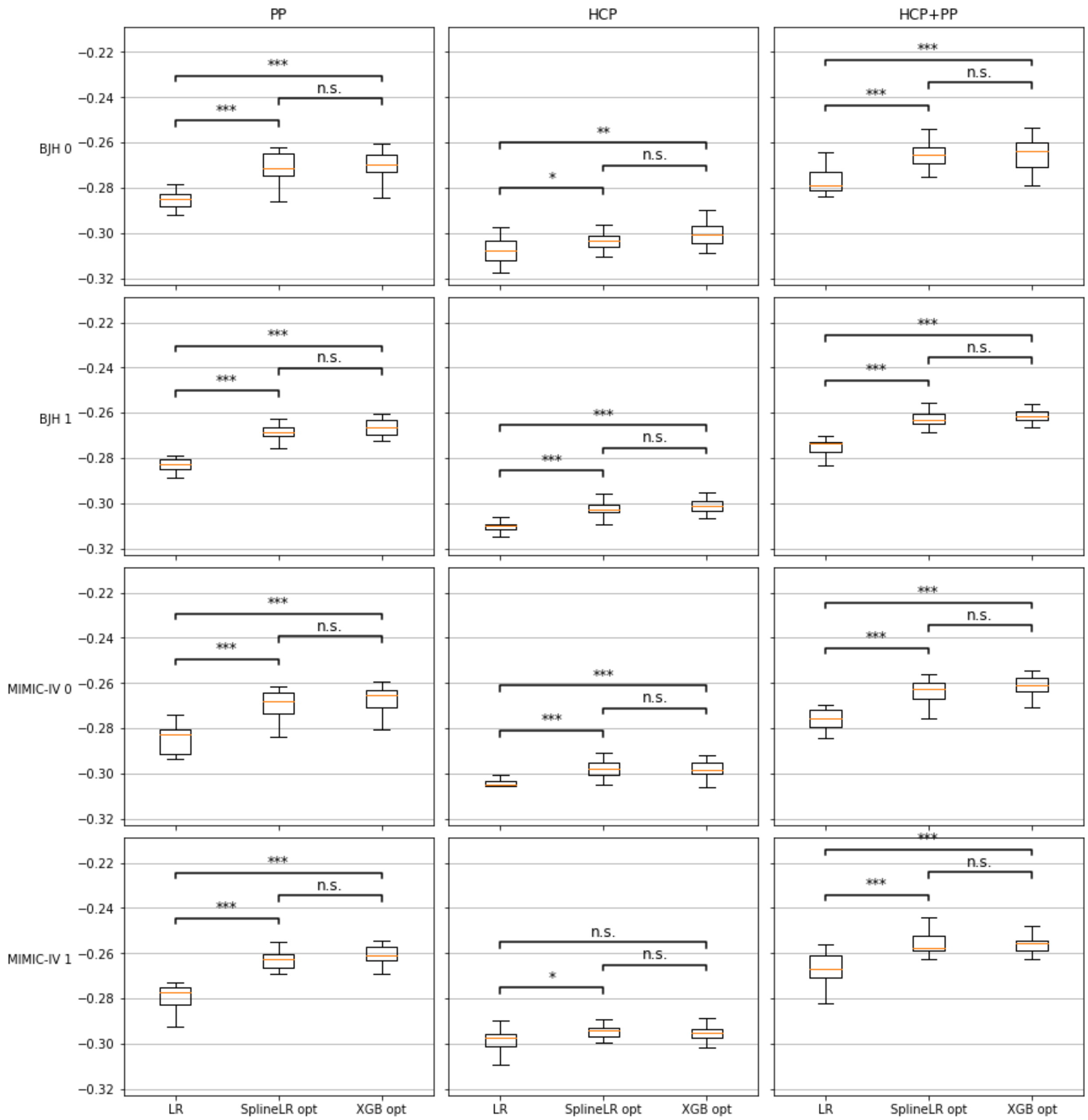


Figure 6. Comparison of Model Types for Internal Performance

Negative log loss (y-axis) was compared across different model types (LR, SplineLR opt, XGB opt), repeated for each site-era (subplot rows) and feature-set combination (subplot columns).

2.4.3.3. Comparison of Feature-Sets for Internal Performance

The internal cross-validation (test) model performance of different feature-sets (HCP, PP, HCP+PP) were compared, repeated for every site, era, and model type combination (**Figure 7**).

In every combination, PP outperformed HCP, and HCP+PP outperformed HCP. In 10 out of 12 combinations, HCP+PP outperformed PP.

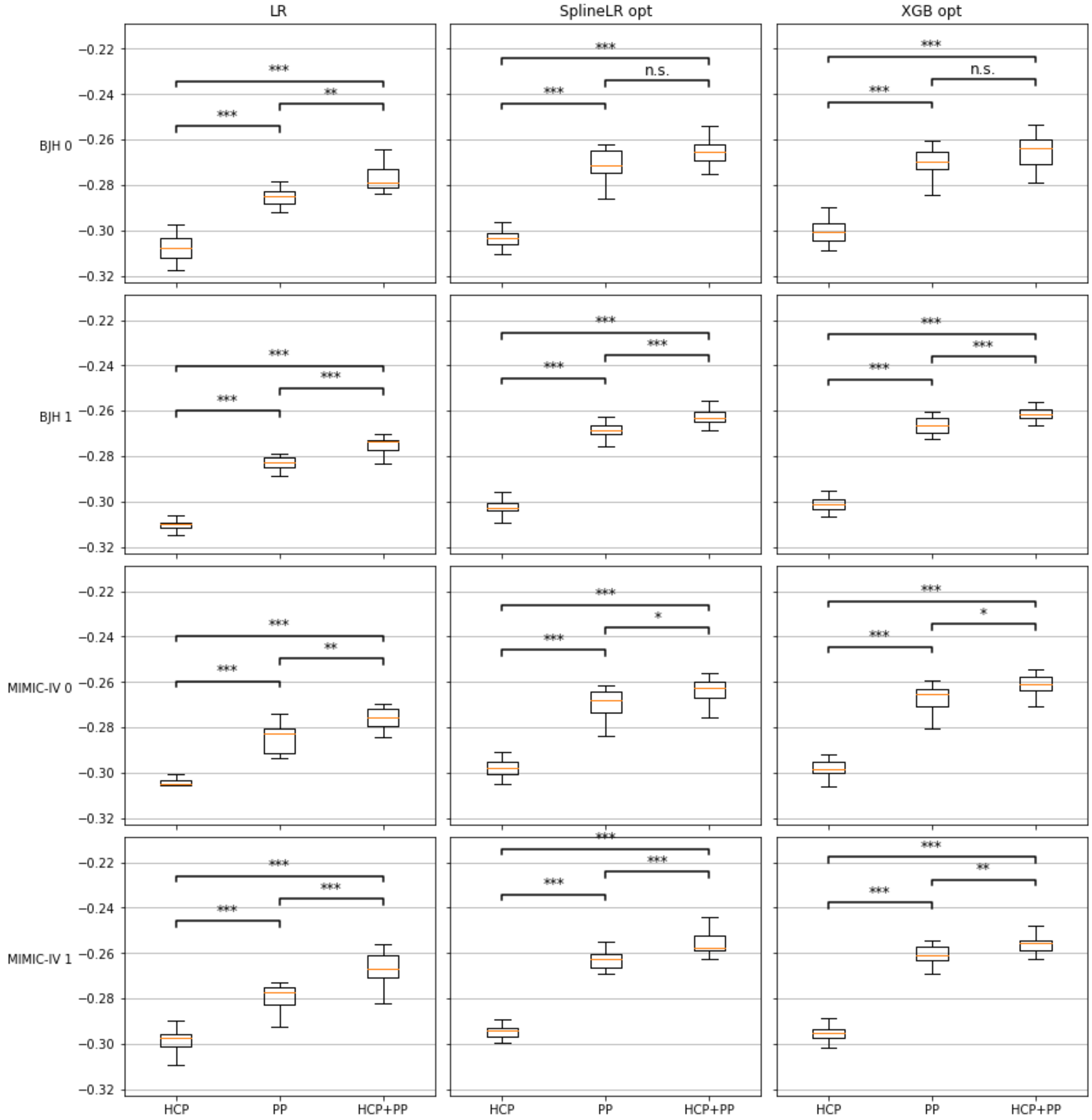


Figure 7. Comparison of Feature-Sets for Internal Performance

Negative log loss (y-axis) was compared across different model types (LR, SplineLR opt, XGB opt), repeated for each site-era (subplot rows) and feature-set combination (subplot columns).

2.4.3.4. Performance Loss on External Use

The internal cross-validation (test) model performance was compared to external (cross-site, cross-time, and cross-site and time) bootstrap performance, stratified by model type, site, era, and feature-set. For LR, significant performance degradation occurred in 7/12 combinations cross-time, in all combinations cross-site, and in all combinations cross-site and time (**Figure 8**). For SplineLR opt, significant performance degradation occurred in 7/12 combinations cross-time, in 11/12 combinations cross-site, and in all combinations cross-site and time (**Figure 9**). For XGB opt, significant performance degradation occurred in 6/12 combinations cross-time, in all combinations cross-site, and in all combinations cross-site and time (**Figure 10**).

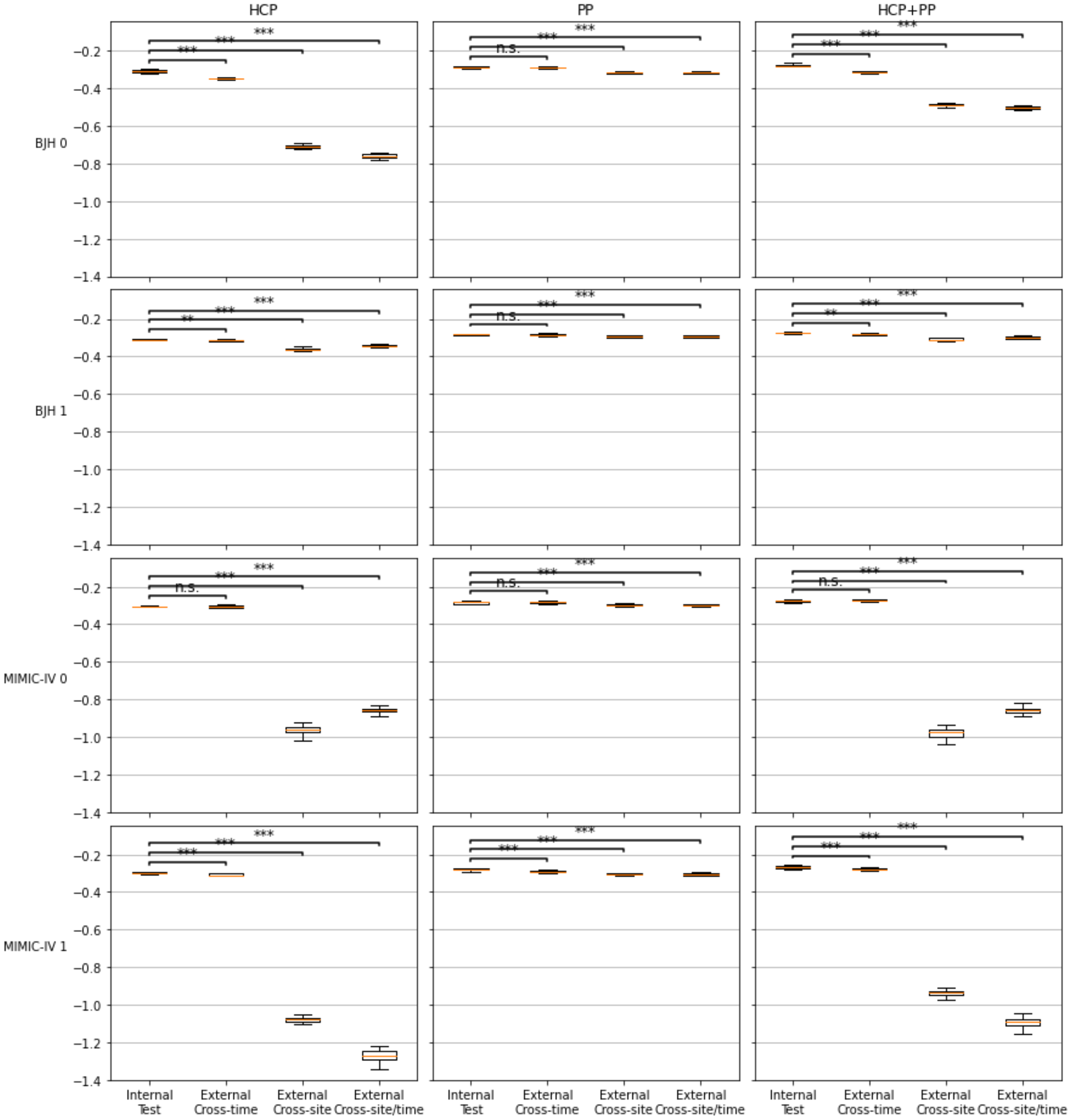


Figure 8. External Validation Performance Loss for LR

Negative log loss (y-axis) was compared between internal and external (cross-time, cross-site, cross-site and time), repeated for each site-era (subplot rows) and feature-set combination (subplot columns).

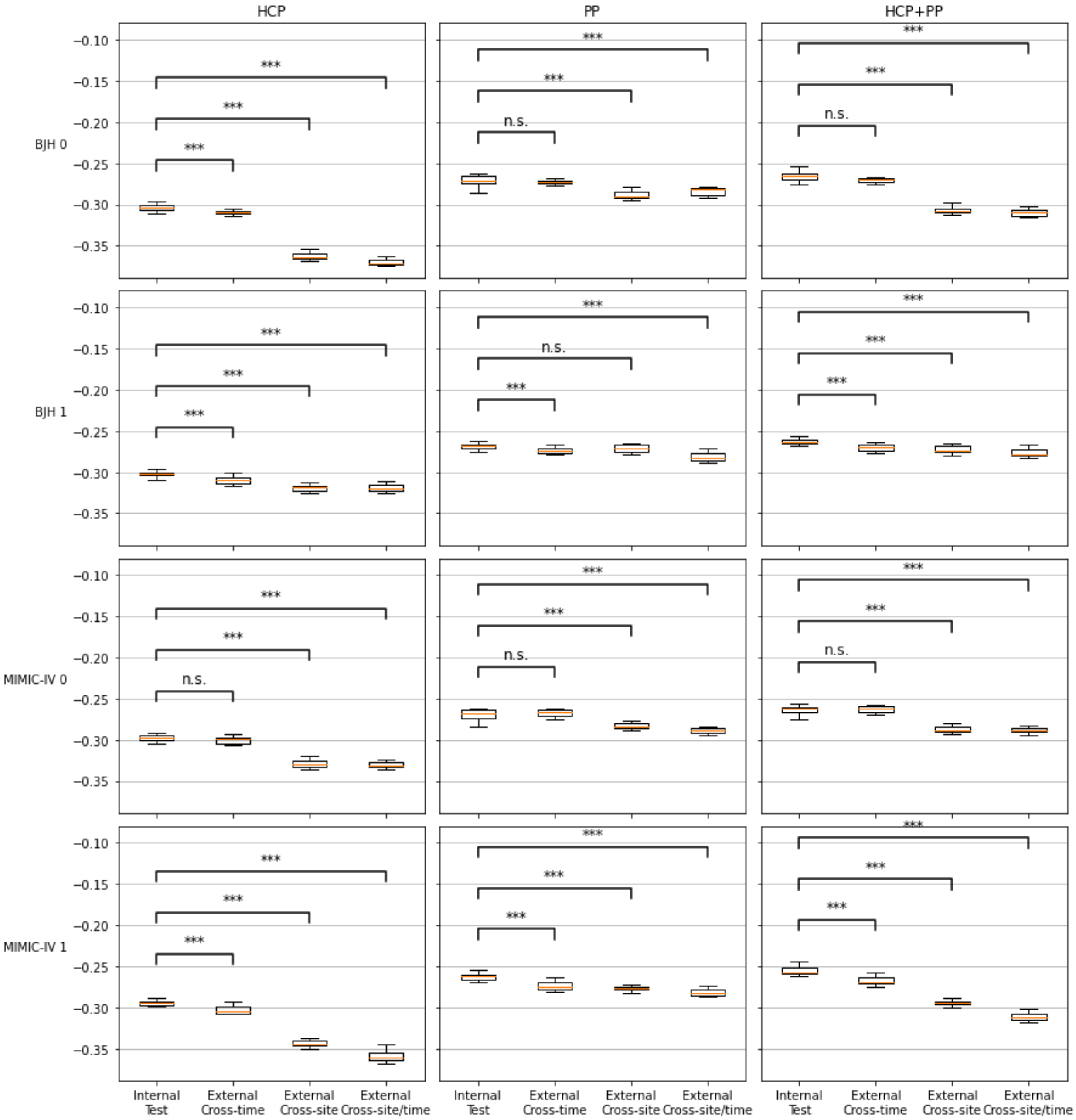


Figure 9. External Validation Performance Loss for SplineLR opt

Negative log loss (y-axis) was compared between internal and external (cross-time, cross-site, cross-site and time), repeated for each site-era (subplot rows) and feature-set combination (subplot columns).

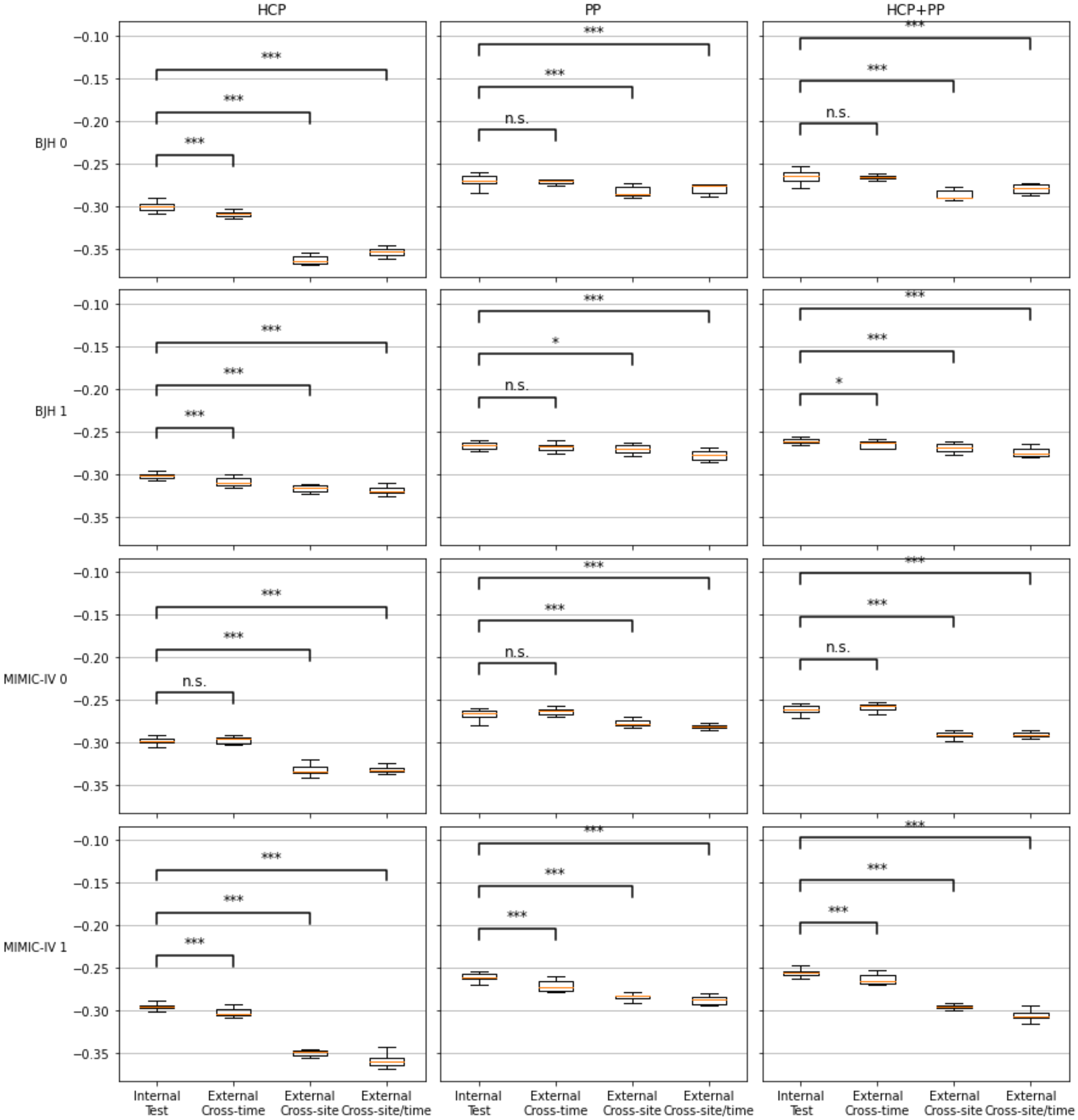


Figure 10. External Validation Performance Loss for XGB opt

Negative log loss (y-axis) was compared between internal and external (cross-time, cross-site, cross-site and time), repeated for each site-era (subplot rows) and feature-set combination (subplot columns).

2.4.3.5. External Use Performance Comparison

The internal and external performances for all combinations of site, era, model type, and feature-set are shown in **Table 5**. For all site-era combinations, the best performing models – internally and externally cross-time – were XGB opt using both HCP and PP features. However, for external cross-site, the best performing models used PP features, with the exception of BJH era 1. The best performing model type cross-site were XGB opt except for in MIMIC-IV era 1 where SplineLR slightly outperformed XGB opt. Of the best performing models cross-site and time, two out of four used PP features, with the other half using HCP+PP; also, three out of four used XGB opt, with one using SplineLR opt.

Table 5. External Use Performance Comparison

Site	Era	Model type	Feature-set	Internal Test (mean ± std)	External Cross-time (mean ± std)	External Cross-site (mean ± std)	External Cross-site & Time (mean ± std)
BJH	0	LR	PP	-0.286 ± 0.005	-0.287 ± 0.003	-0.315 ± 0.004	-0.316 ± 0.005
		LR	HCP	-0.308 ± 0.006	-0.347 ± 0.002	-0.707 ± 0.009	-0.760 ± 0.011
		LR	HCP+PP	-0.278 ± 0.008	-0.311 ± 0.003	-0.488 ± 0.006	-0.501 ± 0.008
		XGB opt	PP	-0.270 ± 0.006	-0.271 ± 0.002	**-0.283 ± 0.006**	-0.280 ± 0.005
		XGB opt	HCP	-0.301 ± 0.005	-0.309 ± 0.003	-0.363 ± 0.005	-0.353 ± 0.004
		XGB opt	HCP+PP	**-0.265 ± 0.007**	**-0.266 ± 0.003**	-0.287 ± 0.005	**-0.279 ± 0.005**
		SplineLR opt	PP	-0.271 ± 0.006	-0.273 ± 0.003	-0.288 ± 0.004	-0.284 ± 0.005
		SplineLR opt	HCP	-0.303 ± 0.004	-0.310 ± 0.002	-0.362 ± 0.004	-0.371 ± 0.004
		SplineLR opt	HCP+PP	-0.267 ± 0.007	-0.270 ± 0.003	-0.307 ± 0.004	-0.309 ± 0.004
BJH	1	LR	PP	-0.283 ± 0.004	-0.285 ± 0.004	-0.294 ± 0.005	-0.296 ± 0.005
		LR	HCP	-0.310 ± 0.002	-0.315 ± 0.005	-0.361 ± 0.006	-0.342 ± 0.005
		LR	HCP+PP	-0.276 ± 0.004	-0.282 ± 0.005	-0.309 ± 0.006	-0.299 ± 0.005
		XGB opt	PP	-0.266 ± 0.004	-0.268 ± 0.004	-0.271 ± 0.005	-0.278 ± 0.006
		XGB opt	HCP	-0.301 ± 0.003	-0.309 ± 0.005	-0.316 ± 0.004	-0.318 ± 0.004
		XGB opt	HCP+PP	**-0.261 ± 0.003**	**-0.265 ± 0.004**	**-0.269 ± 0.005**	**-0.274 ± 0.005**
		SplineLR opt	PP	-0.268 ± 0.003	-0.274 ± 0.004	-0.271 ± 0.005	-0.282 ± 0.005
		SplineLR opt	HCP	-0.302 ± 0.003	-0.310 ± 0.005	-0.319 ± 0.004	-0.319 ± 0.005
		SplineLR opt	HCP+PP	-0.262 ± 0.004	-0.270 ± 0.004	-0.273 ± 0.005	-0.276 ± 0.005
MIMIC-IV	0	LR	PP	-0.285 ± 0.006	-0.283 ± 0.004	-0.300 ± 0.005	-0.299 ± 0.003
		LR	HCP	-0.306 ± 0.004	-0.305 ± 0.004	-0.964 ± 0.026	-0.858 ± 0.016
		LR	HCP+PP	-0.277 ± 0.006	-0.274 ± 0.005	-0.983 ± 0.027	-0.859 ± 0.018
		XGB opt	PP	-0.267 ± 0.005	-0.264 ± 0.004	**-0.277 ± 0.003**	**-0.282 ± 0.003**
		XGB opt	HCP	-0.298 ± 0.004	-0.297 ± 0.004	-0.332 ± 0.006	-0.332 ± 0.003
		XGB opt	HCP+PP	**-0.262 ± 0.005**	**-0.259 ± 0.004**	-0.291 ± 0.004	-0.291 ± 0.003
		SplineLR opt	PP	-0.270 ± 0.007	-0.267 ± 0.004	-0.283 ± 0.004	-0.289 ± 0.003
		SplineLR opt	HCP	-0.298 ± 0.004	-0.301 ± 0.004	-0.329 ± 0.005	-0.331 ± 0.003
		SplineLR opt	HCP+PP	-0.264 ± 0.006	-0.263 ± 0.004	-0.288 ± 0.004	-0.288 ± 0.003
MIMIC-IV	1	LR	PP	-0.279 ± 0.005	-0.290 ± 0.005	-0.306 ± 0.004	-0.306 ± 0.005
		LR	HCP	-0.298 ± 0.005	-0.308 ± 0.005	-1.080 ± 0.021	-1.274 ± 0.035
		LR	HCP+PP	-0.266 ± 0.007	-0.277 ± 0.005	-0.936 ± 0.019	-1.097 ± 0.029
		XGB opt	PP	-0.261 ± 0.005	-0.271 ± 0.006	-0.283 ± 0.004	-0.288 ± 0.005
		XGB opt	HCP	-0.295 ± 0.004	-0.302 ± 0.005	-0.350 ± 0.003	-0.359 ± 0.007
		XGB opt	HCP+PP	**-0.256 ± 0.004**	**-0.263 ± 0.005**	-0.295 ± 0.003	-0.306 ± 0.005
		SplineLR opt	PP	-0.263 ± 0.004	-0.274 ± 0.006	**-0.278 ± 0.003**	**-0.282 ± 0.004**
		SplineLR opt	HCP	-0.294 ± 0.004	-0.303 ± 0.005	-0.344 ± 0.003	-0.359 ± 0.007
		SplineLR opt	HCP+PP	**-0.256 ± 0.005**	-0.268 ± 0.005	-0.295 ± 0.004	-0.312 ± 0.005

Performance – log loss – of mortality prediction models for each combination of site, era, model type, and feature set. Distributions were generated through repeated cross-validation for internal test; and bootstrap for the external generalizability tests.

2.4.3.6. Feature Analysis

Analysis of features focused on XGB opt model type as it was often the best performing model internally and externally. Feature importance was computed using SHAP for each site-era combination, and was ranked based on median SHAP value.[77] The top 20 highest ranked, or most important features are shown in **Table 6**. Of those, 5 were HCP features (shown in red for emphasis): FiO2 count, FiO2 S6 (6th shift) count, O2 flow count, point-of-care glucose count, and total outlier count. The distributions of each of the top 20 features are shown in **0**. Many PP features did not vary drastically across site/eras – for example, the distribution (median, IQR) for Blood Urea Nitrogen (**BUN**) were as follows: BJH 0, 18.5 (12.5 - 30); BJH 1, 18.5 (12.5 - 31.0); MIMIC-IV 0, 19.0 (13.0 - 32.0); MIMIC-IV 1, 18.5 (13.0 - 31.0). However, important HCP features often did vary drastically across sites. For example, FiO2 count were as follows: BJH 0, 1 (0 - 2); BJH 1, 1 (0 - 3); MIMIC-IV 0, 2 (0 - 7); MIMIC-IV, 1 (0 - 7). O2 flow were twice as more frequently documented in BJH than in MIMIC-IV: BJH 0, 10 (0-23); BJH 1, 7 (0-21); MIMIC-IV 0, 3 (0 - 6); MIMIC-IV 1, 4 (0 - 6). Similarly, POC glucose count was much more frequently documented in BJH: BJH 0, 9 (3 – 13); BJH 1, 8 (2 – 12); MIMIC-IV 0, 2 (0 – 4); MIMIC-IV 1, 2 (0 – 4).

Table 6. Feature Importance for XGB opt using HCP+PP

Variable	Variable Type	BJH Era 0	BJH Era 1	MIMIC-IV Era 0	MIMIC-IV Era 1	Median
AGE	PP	0.195	0.264	0.287	0.351	0.276
BUN	PP	0.273	0.265	0.212	0.202	0.239
RDW CV	PP	0.119	0.062	0.214	0.202	0.160
FiO2 (Count)	HCP	0.002	0.148	0.108	0.127	0.117
Anion Gap	PP	0.102	0.101	0.105	0.112	0.103
Albumin	PP	0.149	0.181	0.021	0.037	0.093
PLT	PP	0.098	0.092	0.053	0.089	0.091
Respiratory Rate (q0.75)	PP	0.077	0.103	0.033	0.135	0.090
FiO2 (S6, Count)	HCP	0.001	0.003	0.167	0.164	0.083
Respiratory Rate (q0.25)	PP	0.054	0.102	0.118	0.048	0.078
O2 Flow (Count)	HCP	0.167	0.112	0.034	0.015	0.073
WBC	PP	0.021	0.067	0.066	0.109	0.066
Lactate	PP	0.110	0.059	0.052	0.072	0.065
PTT	PP	0.065	0.035	0.104	0.064	0.065
Temperature	PP	0.054	0.066	0.033	0.102	0.060
Alkaline Phosphatase	PP	0.044	0.094	0.049	0.066	0.058
Arterial pH	PP	0.103	0.090	0.018	0.018	0.054
POC Glucose (Count)	HCP	0.025	0.037	0.059	0.084	0.048
Heart Rate (q0.50)	PP	0.017	0.046	0.049	0.046	0.046
Total Outlier Count	HCP	0.034	0.060	0.027	0.057	0.046

Top 20 most important features based on median SHAP value across the four site-era combinations of the XGB opt model using both HCP and PP features. HCP features are highlighted in red for emphasis.

2.4.4. Discussion

In this study, we explored the impact of features heavily influenced by healthcare processes, such as frequency of lab tests, on clinical prediction model performance and generalizability. Compared to prior work, the types of HCP variables, modes of generalizability, and model types were all expanded to provide a more comprehensive view on the impact of HCP features on generalizability. What we find is that when performing internal validation, the addition of HCP features significantly improved model performance. Applied cross-time, the performance improvement (compared to models using only PP features) persisted. However, applied cross-site, HCP+PP models were often less performant than models using PP only. Investigation of features show that important HCP variables often differ drastically between sites which are unlikely to be due to underlying patient population pathophysiology, but rather, cultural and procedural differences between sites.

As expected, vanilla logistic regression performed significantly worse than all other compared model types. However, augmenting with relatively simple feature transformation and optimization of hyperparameters – in our case, basis spline feature expansion and optimization of regularization parameter – yielded performance that was just as good as the more complex Gradient Boosted Decision Tree (**GBDT**) models – in our case, XGB. This result emphasizes the value of incorporating non-linearity to the CPM development process, and implies diminishing returns with added complexity. We did not compare against deep neural networks or variants thereof including recurrent variants due to concerns regarding interpretability. However, there are novel neural network architectures that trade performance for interpretability, which if seen wide adoption, should also be assessed as well.[78]

Models only using HCP features performed poorly compared to those using PP or both. The implication is that while HCP features do provide additional information to the model, HCP by themselves are insufficient, especially compared to models using PP features. Further, given the non-generalizability of models relying on HCP features cross-site, they should be given special attention – either dropping them from the model or fine-tuning when applying cross-site.

There is a growing set of literature investigating the relationship and disconnect between what's present in EHR data and the true underlying patient physiology, and more recently, the impact of this phenomenon on clinical prediction modeling. However, there is a lack of terminological or conceptual unification, resulting in what appears to be scattered or piecemeal analyses. Much of this work falls under investigating bias in EHR data.[71, 79] Some focus narrowly on missing data imputation or measurement indicators.[73, 80, 81] Others have proposed frameworks for understanding the phenomenon – we followed the Hripscsak and Albers framework as interpreted by Agniel et al. in which HCP represents the recording process which distorts the EHR data representation of the true underlying patient state.[66, 71] Under this framework, when a lab test is ordered would be considered subjective, whereas the lab test result would be considered objective. However, others have a more conservative definition for objective data, and argue that the vast majority of EHR data is subjective as it is "clinician-initiated," meaning that only measurements that do not require clinician intervention like routine lab tests or telemetry data are unfiltered or unbiased. In summary, many researchers have been tackling the issue of EHR data being a filtered view of the patients' true state, some of whom have developed ad-hoc or piecemeal theories that conflict with one another on certain points. We believe that a consensus framework and a shared vocabulary will facilitate understanding and accelerate research in this field.

The best performing model and feature-set combination based on the internal cross-validation often did not generalize well (cross-site). In other words, the typical optimal model selection procedure resulted in a suboptimal model when applied externally. Ideally, external sites would fully re-train or fine-tune the model on their own historic data to prevent significant performance degradation. However, it's often unknown where the model will be considered for implementation, and even then, it may be prohibitive to acquire data from all those organizations. Thus, methods to a priori determine robustness or resilience against non-generalizability due to drift/shift would be very beneficial for building generalizable models. There is ongoing research in this area, often under the umbrella of machine learning safety, on building models that are robust to dataset shift.[41] While we cannot offer a domain-agnostic advice, our results suggest that for EHR data specifically, if generalizability is a priority, HCP features should be used with care or not used at all.

The external cross-site and cross-site/time performance loss for models using HCP features was significantly higher for models trained in BJH era 0 than BJH era 1 across all model types (**Figure 8, Figure 9, Figure 10**). Notably, BJH era 1 covers a system-wide enterprise EHR transition on June 2018. We also observed that the EHR transition resulted in many changes to the EHR data including extreme changes in frequency for certain measurements. Thus, we hypothesize that the EHR transition represents a drastic shift in the healthcare process, thus emulating multi-site data, thereby improving generalizability, and modulating cross-site performance loss.

2.4.5. Conclusion

Addition of features based on variables heavily influenced by the healthcare process, such as frequency of lab tests, appear to improve clinical prediction model performance, but those models then have limited generalizability or lower performance when applied on external sites. Thus, healthcare process features must be used with care, or not used at all.

2.5. Study 2: Respiratory Support Status: Development of Standards

Identifying the lack of standards as a potential driver of CPM non-transportability and identifying respiratory support methods as a domain of EHR data lacking in standards, this study proposes and evaluates a novel classification standard for respiratory methods as well as accompanying heuristics for mining respiratory support episodes.

2.5.1. Introduction

Managing respiratory status and providing appropriate respiratory support to prevent or mitigate hypoxemia is a critical aspect of clinical management, especially for patients suffering from respiratory conditions such as COVID-19.[82] Leveraging respiratory support information not only provides a more complete clinical picture, but also can be used in downstream analyses such as sub-phenotyping or predictive modeling as a source of features or to identify endpoints.[83, 84] However, generating generalizable conclusions from respiratory support information is difficult due to the heterogeneity of respiratory support methods and settings, the lack of standardized representations, and the poor quality of data regarding respiratory support.[85]

Standardization of patient data representation has accelerated knowledge discovery, replication of results, and translation into practice. While many parts of healthcare information have been standardized – such as the ICD codes for diagnoses, LOINC for measurements and observations, or the all-encompassing Unified Medical Language System (**UMLS**) – much of the patient information in EHR data, such as clinical cultures and respiratory support methods, remains unmapped.[86, 87] Standardization of respiratory support is made challenging due to the inherent diversity of methods, especially when considering methods designed for subpopulations, variations, and modifiable settings. Worse yet, the lack of a widely adopted standardization schema has resulted in the usage of highly heterogeneous terms in literature and practice, even for identical concepts – for example, heated and humidified high-flow nasal cannula has been described as high flow nasal cannula, high humidity nasal cannula, high flow nasal oxygen therapy, or referred to by brand names such as Airvo™ or Optiflow™.[88-90] Prior efforts to standardize respiratory support terms resulted in high granularity, which while necessary for coverage, can be excessive for downstream tasks. For example, the "Respiratory Therapy" concept in Systematized Nomenclature of Medicine (**SNOMED-CT**) contains 14 children, of which "Oxygen Therapy" itself contains 14 children.[91] Thus, there is a need for a pragmatic and parsimonious standardization of respiratory support terms.

Beyond the lack of standardization, extraction of respiratory support status from EHR data is made challenging due to scattered, incomplete, and often contradictory documentation which necessitates the usage of auxiliary documentation and heuristics to determine respiratory support status.[74, 92, 93] Prior heuristic development work, however, are dataset-specific and focus on endotracheal intubation thus fail to address other strata of respiratory support.[74, 92]

Therefore, the objective of the study was to: 1) propose a preliminary, parsimonious, and pragmatic terminology system for respiratory support stratified by severity of hypoxemia; 2) develop (meta-)heuristics for the construction of respiratory support episodes from raw and heterogeneous EHR data; and 3) evaluate the terminology system and heuristics by measuring its impact on 30-day mortality prediction through assessment of feature importance and feature ablation studies.

2.5.2. Methods

2.5.2.1. Study Design, Data Sources, and Population

All patients ≥ 18 years of age admitted to all hospitals within a large Midwestern healthcare system serving the metropolitan St. Louis, mid-Missouri, and southern Illinois regions between 3/1/20 and 4/1/21 were eligible for inclusion. Patients were included if they had a COVID-19 Polymerase Chain Reaction (**PCR**) or antigen test, positive or negative, within 14 days prior to or 7 days after hospital admission. Patients were excluded if they had a hospital length of stay (**LOS**) < 24 hours to allow for a sufficient observation window for the predictive modeling study. Patients without associated demographics, comorbidities, or location data were excluded. Only patients with at least 5 heart rate and 5 SpO2 measurements during the first 24 hours of hospital arrival were included. EHR data, including demographics, vital signs, lab results, flowsheet entries, etc., were extracted for all included subjects. This project was approved with a waiver of informed consent by the Washington University in St. Louis Institutional Review Board.

2.5.2.2. Classification System of Respiratory Support Methods

To facilitate generalizability and reproducibility of studies leveraging respiratory support information, we propose the following terminologies, in increasing severity: Low Flow Oxygen Therapy (**LFOT**), High Flow Oxygen Therapy (**HFOT**), Non-Invasive Mechanical Ventilation (**NIMV**), Invasive Mechanical Ventilation (**IMV**), and ExtraCorporeal Membrane Oxygenation (**ECMO**).[74, 94]

2.5.2.3. Meta-heuristics for Identification of Respiratory Support Episodes

The authors of MIMIC-III, in addition to the data, also published code for data processing and analysis to expedite and encourage collaborative research.[74] Among the SQL scripts in their GitHub repository is one for calculating mechanical ventilation duration. Essentially, their logic was to chain together proximal pieces of documentation that are indicative of mechanical ventilation to form episodes with start and end times. Their heuristic has since been used successfully by researchers using the MIMIC-III dataset.[93] To generalize the heuristic for other forms of respiratory support beyond mechanical ventilation, and for other datasets beyond MIMIC-III, we developed a generalized version of the MIMIC-III heuristic – a "meta-heuristic" – to guide the development of heuristics for the assembly of respiratory support episodes as follows:

1. Define two parameters:
    a. MIN_DURATION for the minimum episode duration
    b. EXTENSION_TOLERANCE for the maximum allowable time gap between documentation for the formation of episodes
2. Identify timestamped documentation that are indicative of the presence of respiratory support
3. Link consecutive documentation occurring within EXTENSION_TOLERANCE into episodes
4. Discard any episodes with duration less than or equal to MIN_DURATION

Next, as we conceived of the respiratory support methods as being mutually exclusive, respiratory support episodes are "flattened" into a single timeline such that at any given time, a patient is on either no respiratory support or a single respiratory support method, by giving higher severity methods priority. Finally, the respiratory support trajectories are "repaired" such that gaps between episodes with a duration less than EXTENSION_TOLERANCE are filled by extending the preceding episode. Also, gaps at the beginning and end of the patient stay (between encounter start time and first respiratory support episode start time, and between last respiratory support episode end time and encounter end time), if less than MIN_DURATION, are filled by extending the first or last episode, respectively.

2.5.2.4. Evaluation through In-hospital Mortality Prediction

A predictive modeling study was designed to evaluate the utility of the respiratory support information extracted through our heuristics on downstream analyses. The task was to predict in-hospital mortality within 30 days, at 24 hours after hospital arrival for a COVID-19-tested, adult cohort presenting to an ED. 121 **baseline** features were generated from demographic, laboratory, vital sign, and other clinical data extracted from the EHR. For most numeric measurements, the median value during the observation window was extracted, but for frequent measurements such as heart rate, other distributional statistics (25th quantile, 75th quantile, and interquartile range) were also extracted.

10 additional, respiratory-support-derived features were generated using the **proposed** classification schema and heuristics which included duration of respiratory support per type, and the last respiratory support during the observation period. For comparison, we identified a small set of measurements **related** to respiratory status: Fraction of inspired Oxygen (**FiO2**) and oxygen flow rate. We also extracted the EHR-native representation of respiratory support called oxygen delivery method (**O2 Del Method**) which included ETT, CPAP, T-Piece, etc. Lastly, we also considered a set of features based on the proposed classification, but using the **raw** time-stamped data prior to assembly into episodes. The feature sets including explicit respiratory support information – O2 Del Method, Raw, and Proposed – all also include “Baseline” and “Related” features (**Appendix 8, Appendix 9**).

The compared algorithms were logistic regression (**LogReg**) and XGBoost (**XGB**). For LogReg, features were standardized and mean-imputed whereas for XGB, features were left as is. Hyperparameters such as regularization strength were optimized for log loss using the baseline features through 1,000-iteration, 4-fold cross-validation (**Appendix 10**). Once optimal hyperparameters were identified, 5 replicates of 2-fold cross-validation was performed to generate a distribution of performance metrics: Area Under Receiver Operating Characteristic curve (**AUROC**), Area Under Precision Recall Curve (**AUPRC**), and negative log loss.[76] The distributions of performance metrics were compared using the Wilcoxon signed-rank test (two-way, paired).[95] Feature importance was quantified using SHapley Additive exPlanations (**SHAP**) values for the XGBoost model, and coefficient values for the logistic regression model , both of which were aggregated over 100 bootstrap samples.[77, 96]

2.5.2.5. Statistical Analysis

Variables were summarized using frequencies and proportions for categorical data and medians and interquartile ranges or means and standard deviations for continuous data. Statistical comparisons were performed using the Chi-square and Mann-Whitney U tests where appropriate unless specified otherwise. A p-value < 0.01 was considered statistically significant. All resampling analyses, cross-validation and bootstrap, were performed using a fixed seed. All analysis and figure generation were performed with Python version 3.7.1 (Python Software Foundation, Beaverton, OR) using the following packages: scipy, numpy, pandas, matplotlib, sklearn, xgboost, and shap.[77, 97-102]

2.5.3. Results

The severity-stratified respiratory support methods and the documentation serving as evidence for each method are listed in **Table 7**. Examples of the full heuristic application process, along with documentation germane to respiratory status, can be found in **Figure 11. Figure 12** shows respiratory support utilization over time, with patients temporally aligned at ED arrival.

Table 7. Respiratory support terminologies and heuristics

Term	**Institution-specific Evidence**
Low Flow Oxygen Therapy (LFOT)	Oxygen delivery method documentation: nasal cannula, non-rebreather mask, simple mask, venturi mask, aerosol mask, face tent. Oxygen flow rate ≤ 15 L / min
High Flow Oxygen Therapy (HFOT)	Oxygen delivery method documentation: high flow nasal cannula, high humidity nasal cannula, optiflow Oxygen flow rate > 15 L / min
Non-Invasive Mechanical Ventilation (NIMV)	Documentation of “NPPV Status” as “In Use” in the “Adult NPPV/NIV” flowsheet
Invasive Mechanical Ventilation (IMV)	Documentation of “Vent Status” as “In Use” in the “Ventilator Documentation” flowsheet
ExtraCorporeal Membrane Oxygenation (ECMO)	Documentation of “Pump Flow (L/min)” or “ECMO Pump Speed (RPM)” in the ECMO or VAD flowsheets

Respiratory support method categories as well as the accompanying evidence of respiratory support usage in EHR data to be used in the heuristic for mining respiratory support episodes.

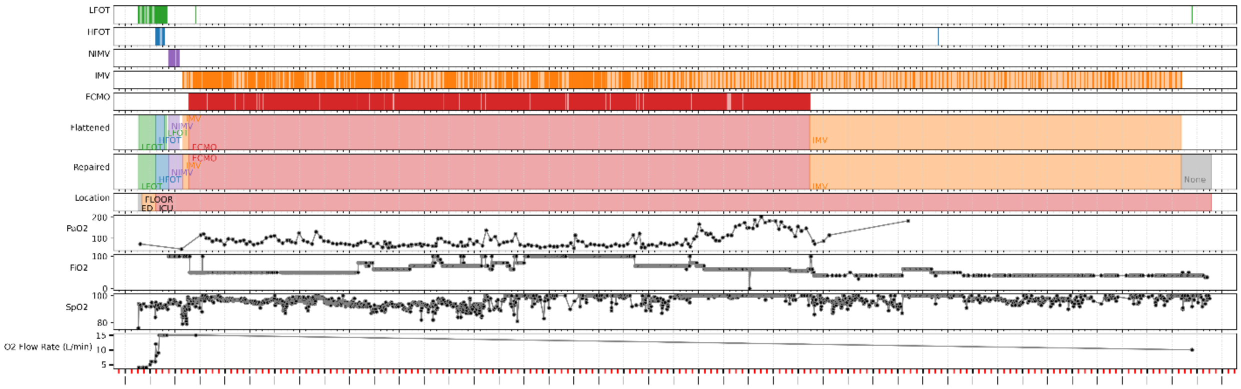


Figure 11. Respiratory Support Trajectory Example

This plot demonstrates the application of the respiratory support trajectory heuristics on a full, single patient encounter using data elements from Table 1, MIN_DURATION of 6 hours, and EXTENSION_TOLERANCE of 24 hours. The x-axis indicates time with each black tick indicating 24 hours and red tick indicating 6 hours. The top 5 subplots each pertain to a single respiratory support method where vertical lines indicate the times at which pieces of documentation serving as evidence for respiratory support were documented. The individual sub-trajectories are all merged into a single timeline as shown in the "flattened" subplot, after which it is repaired as according to the heuristic. The subplots below the "repaired" subplot provide context for the patient, showing patient location and measurements pertaining respiratory status.

Abbreviations: ED, emergency department; ICU, intensive care unit; LFOT, low flow oxygen therapy; HFOT, high flow oxygen therapy; NIMV, non-invasive mechanical ventilation; IMV, invasive mechanical ventilation; ECMO, extracorporeal membrane oxygenation.

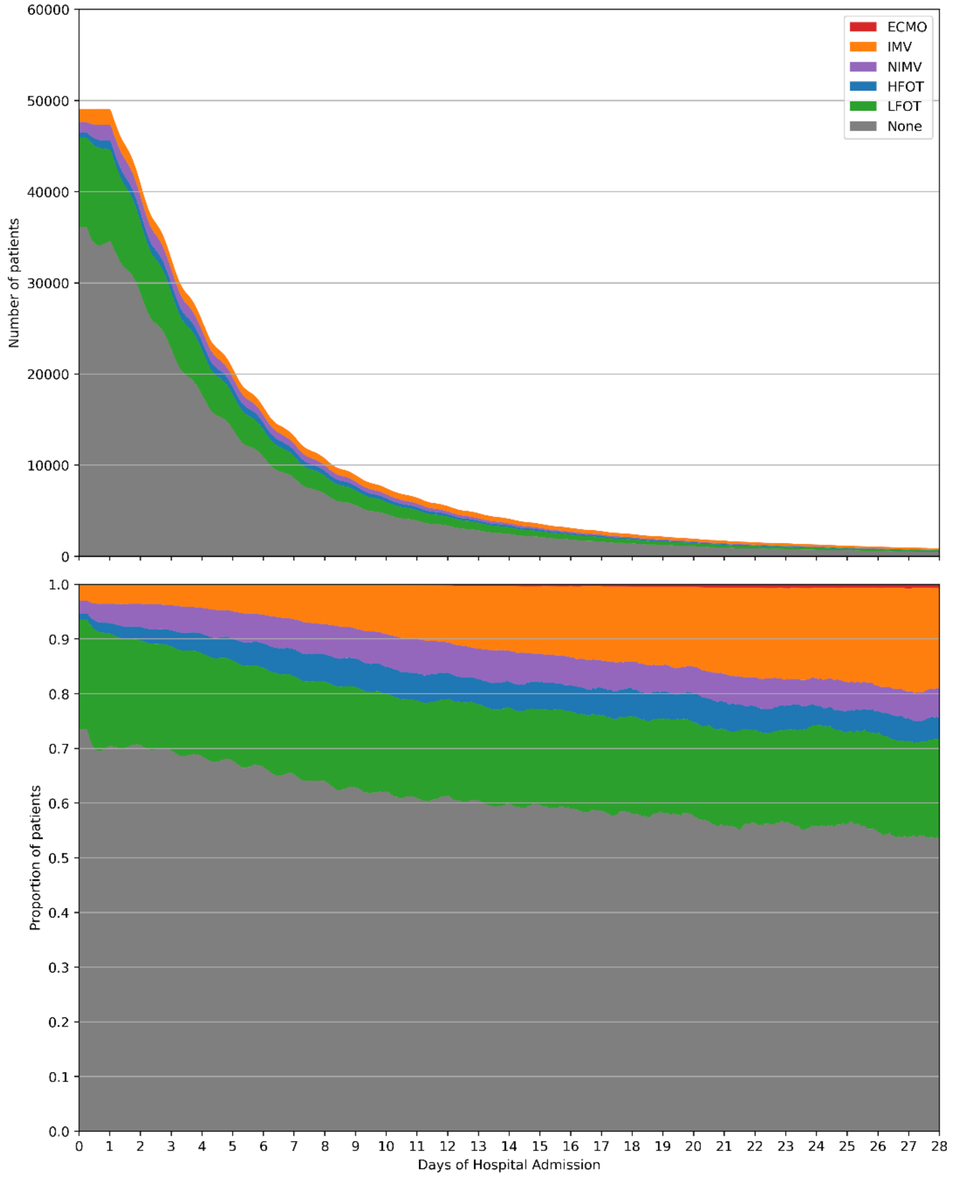


Figure 12. Respiratory Support Utilization Aligned at Arrival

All patients were aligned at ED arrival, and their usage of respiratory support was plotted for the first 4 weeks. The top subplot shows the total number of patients utilizing each respiratory support method, whereas the bottom subplot shows the proportion. As expected, patients who have been in the hospital longer are more likely to be on higher levels of respiratory support.

Abbreviations: LFOT, low flow oxygen therapy; HFOT, high flow oxygen therapy; NIMV, non-invasive mechanical ventilation; IMV, invasive mechanical ventilation; ECMO, extracorporeal membrane oxygenation.

Cohort characteristics and outcomes for the patient population used in this study can be seen in **Table 8**. During the study period there were 45,908 hospitalizations lasting at least 24 hours available for analysis. Of these, 1,601 (3.5%) experienced in-hospital death within 30 days. Non-survivors were older, more likely to be male, more likely to be COVID-19 positive, and have a longer length of stay (**Table 8**).

Table 8. Cohort characteristics

		Outcome = in–hospital mortality within 30 days of index time		
Variable	**Total (n = 45,908)**	**Yes (n = 1,601 [3.5%])**	**No (n = 44,307 [96.5%])**	**p**[a]
Age (years), median (IQR)	64.0 (51.0 – 76.0)	73.0 (62.0 – 83.0)	64.0 (50.0 – 76.0)	< 0.01 *
Male, n (%)	22,638 (49.3%)	889 (55.5%)	21,749 (49.1%)	< 0.01 *
Race, n (%)				< 0.01 *
White	28,032 (61.1%)	933 (58.3%)	27,099 (61.2%)	0.021
Black	16,706 (36.4%)	576 (36.0%)	16,130 (36.4%)	0.747
Asian	340 (0.7%)	17 (1.1%)	323 (0.7%)	0.168
Other/unknown	830 (1.8%)	75 (4.7%)	755 (1.7%)	< 0.01 *
BMI, median (IQR)	27.5 (23.2 – 33.4)	27.1 (22.7 – 32.2)	27.5 (23.2 – 33.4)	< 0.01 *
COVID–19 positive, n (%)	8,332 (18.1%)	502 (31.4%)	7,830 (17.7%)	< 0.01 *
Respiratory support duration during observation window (hours), mean ± std				
None	16.7 ± 10.4	7.5 ± 10.2	17.0 ± 10.2	< 0.01 *
LFOT	5.31 ± 9.16	7.36 ± 10.05	5.24 ± 9.12	< 0.01 *
HFOT	0.39 ± 2.70	1.86 ± 5.73	0.34 ± 2.50	< 0.01 *
NIMV	0.76 ± 3.74	1.47 ± 5.23	0.73 ± 3.68	< 0.01 *
IMV	0.85 ± 4.21	5.79 ± 9.73	0.67 ± 3.75	< 0.01 *
ICU transfer, n (%)	10,311 (22.5%)	1,321 (82.5%)	8,990 (20.3%)	< 0.01 *
Total LOS (hours), median (IQR)	99.7 (59.0 – 172.7)	166.8 (83.4 – 313.6)	98.6 (58.2 – 170.1)	< 0.01 *
In–hospital mortality, n (%)	1,682 (3.7%)	1,601 (100.0%)	81 (0.2%)	< 0.01 *

Abbreviations: IQR, interquartile range; BMI, body mass index; COVID–19, Coronavirus disease 2019; std, standard deviation; LFOT, low flow oxygen therapy; HFOT, high flow oxygen therapy; NIMV, non–invasive mechanical ventilation; IMV, invasive mechanical ventilation; ECMO, extracorporeal membrane oxygenation; ICU, intensive care unit; LOS, length of stay.

[a] Comparison of variables between those with and without the primary outcome of 30–day in–hospital mortality was performed using Mann–Whitney U test for continuous variables, and $\chi 2$ for categorical variables. Statistical significance, $p < 0.01$, is indicated by *.

The optimized hyperparameters (**Appendix 10**) were used for generating distributions of performance metrics through repeated cross-validation (**Figure 13**). For both XGB and LogReg, the addition of “Related” features significantly improved on “Baseline,” and the addition of “O2 Del Method,” “Raw,” or “Proposed” features improved on “Related” across all three metrics: AUROC, AUPRC, and negative log loss (**Figure 13, Appendix 11, Appendix 12**). However, “O2 Del Method,” “Raw,” and “Proposed” rarely differed significantly, and when they did, the differences were very small as was the case for LogReg AUROC between “O2 Del Method” and “Proposed” (0.887 [0.884 - 0.890] and 0.887 [0.885 - 0.891], $p < 0.01$, **Appendix 12**).

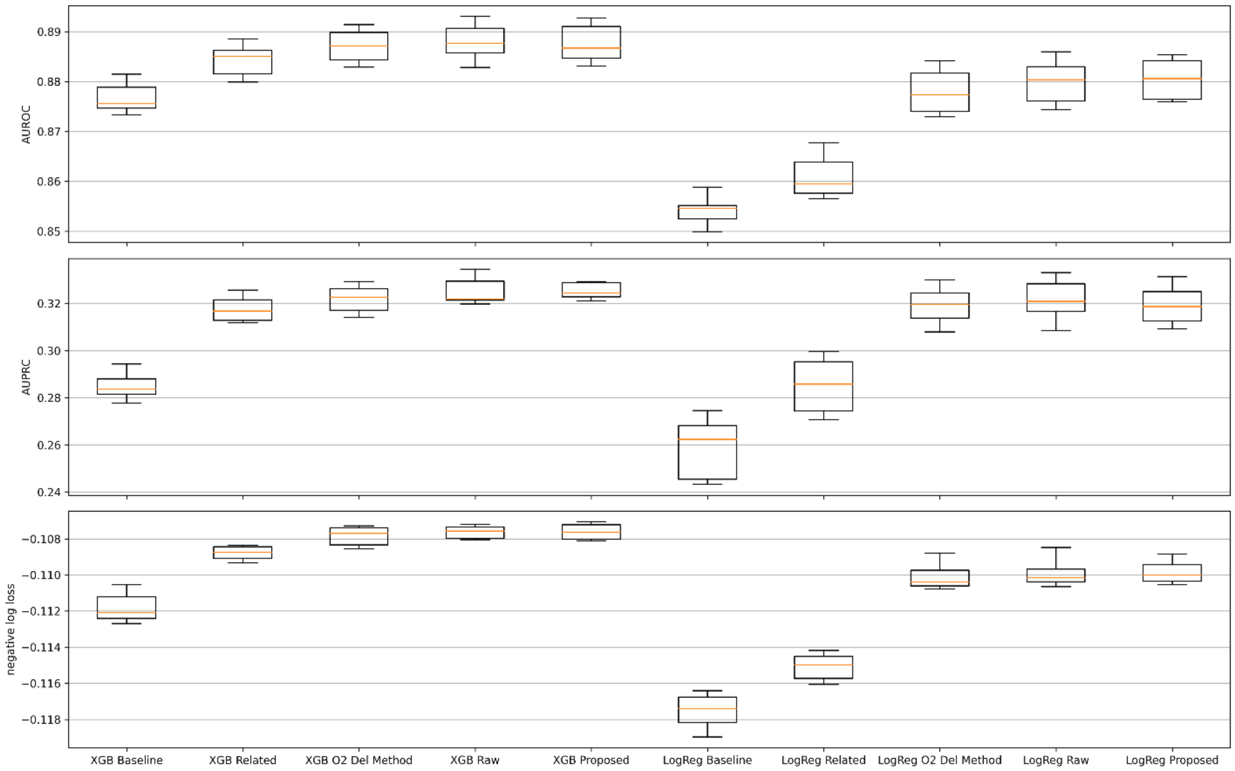


Figure 13. Mortality Prediction Performance Comparison

Comparison of in-hospital mortality prediction performance for LogReg and XGB models with varying sets of features from 5-repeat, 2-fold cross-validation. "Baseline" includes demographics, common lab results, and vital signs from the EHR data. "Related" also includes O2 flow rate and fraction of inspired oxygen. In addition, "O2 Del Method" includes the EHR-native representation of respiratory support status, "Raw" includes data from the proposed approach prior to assembly into episodes, and "Proposed" includes features derived from respiratory support episodes based on the proposed approach. The center horizontal line represents median, box represents the interquartile range between 25th and 75th percentiles, and whiskers represent 2.5th and 97.5th percentiles.

Abbreviations: LogReg, logistic regression; XGB, extreme gradient boosted trees model; AUROC, area under receiver operating characteristic curve; AUPRC, area under precision recall curve.

Six of the top twenty most impactful features for the LogReg model were respiratory-support-derived features, including: last respiratory support, IMV and LFOT duration (**Figure 14**). For the XGB model, respiratory-support-derived features ranked 10th (last respiratory support, IMV) and 17th (LFOT duration) (**Figure 15**).

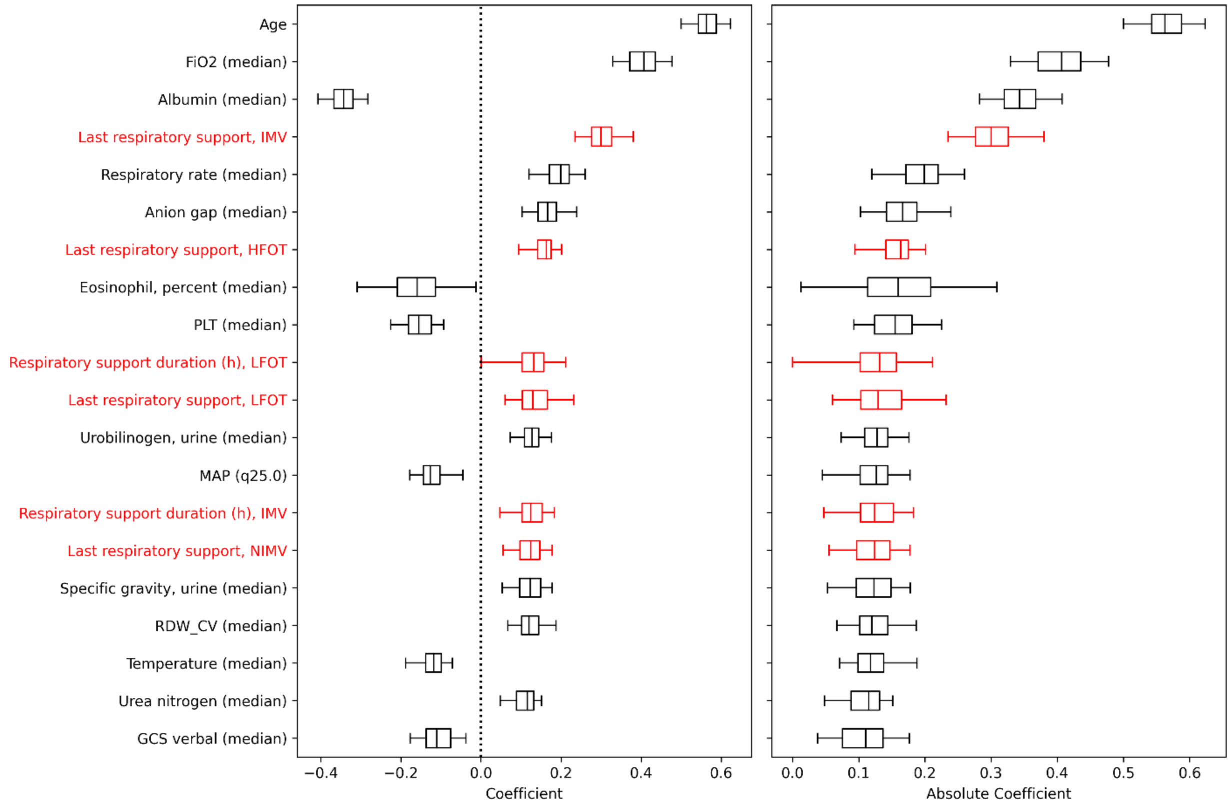


Figure 14. Logistic Regression Feature Importance

The left subplot shows the top 20 most important features in the logistic regression model based on coefficient values aggregated over 100 bootstrap samples, and the right subplot shows the absolute coefficient values. For each feature, the center vertical line represents median, box represents the interquartile range between 25th and 75th percentiles, and whiskers represent 2.5th and 97.5th percentiles. Features based on respiratory support information are colored/shaded in red.

Abbreviations: FiO2, fraction of inspired oxygen, IMV, invasive mechanical ventilation; HFOT, high flow oxygen therapy; PLT, platelet count; LFOT, low flow oxygen therapy; MAP, mean arterial pressure; NIMV, non-invasive mechanical ventilation; RDW_CV, red blood cell distribution width coefficient of variation; GCS, Glasgow Coma Scale.

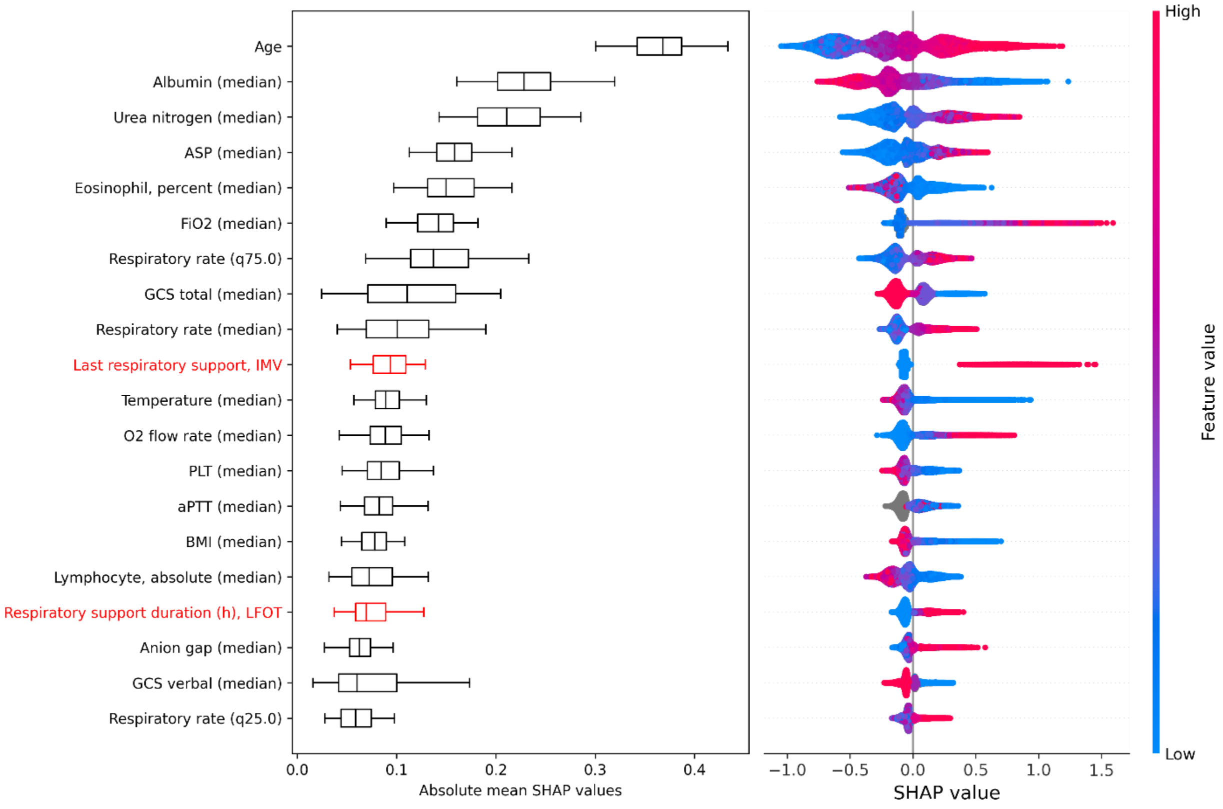


Figure 15. XGBoost SHAP Feature Importance

The left subplot shows the top 20 most important features in the XGB model based on absolute mean SHAP values aggregated over 100 bootstrap samples. The right subplot shows the individual SHAP value for the same top 20 features, for all encounters in the full dataset.

For each feature, the center vertical line represents median, box represents the interquartile range between 25th and 75th percentiles, and whiskers represent 2.5th and 97.5th percentiles. Features based on respiratory support information are colored/shaded in red.

Abbreviations: ASP, aspartate aminotransferase; FiO2, fraction of inspired oxygen; GCS, Glasgow Coma Scale; IMV, Invasive Mechanical Ventilation; PLT, platelet count; aPTT, activated partial thromboplastin time; BMI, body mass index; LFOT, low flow oxygen therapy.

2.5.4. Discussion

In this study, we 1) propose a preliminary, parsimonious, and pragmatic terminology system for respiratory support methods, 2) develop (meta-)heuristics for extraction of respiratory support information from EHR data, and 3) investigate the utility of the respiratory support information extracted through the proposed terminology system and heuristics via a mortality prediction study in a COVID-19-tested, ED-admit, adult cohort. The developed heuristic was successfully applied to EHR data to extract respiratory support episodes, which were then used for in-hospital mortality prediction as features, which were found to be among the most important features for both LogReg and XGB models.

Compared to models using demographics and commonly documented lab results and vital signs, the addition of respiratory-support-related information, FiO2 and O2 flow rate, significantly improved prediction performance for both XGB and LogReg across all measured performance metrics. Moreover, the additional inclusion of explicit respiratory support information further improved performance significantly, again for both model types and across all metrics.

Because there are no other dataset-agnostic, full-severity-spanning classification and heuristic system for respiratory support information found in literature, we compared our proposed approach against two other methods of explicit respiratory status representation: "O2 Del Method" which used the EHR-native representation and "Raw" which uses the data elements for the proposed approach but prior to assembly into episodes (**Appendix 9, Appendix 11**). The proposed representation had commensurate performance to alternate representations of explicit respiratory support information, despite loss of both conceptual and temporal granularity resulting from the aggregation of heterogeneous timestamped raw data into the more human-understandable format of encounter-spanning series of episodes (**Figure 11**).

The increase in model performance associated with the addition of explicit respiratory support information in XGB was less than that of LogReg. We hypothesize that more complex models are able to infer respiratory support status or reconstitute information contained in respiratory status based on other features.

For the models using features from the proposed approach, IMV status and duration was an important predictor which is unsurprising – patients who are intubated are known to have higher rates of in-hospital mortality.[103] This result simply underscores the importance of leveraging respiratory support information, especially those of high severity, for understanding patient status. However, even features based on low flow oxygen therapy status were among the most important features for both the XGB and LogReg models, indicating that lower severity respiratory support information is also critical for developing a complete clinical picture of patients.

In this study, respiratory support information was used as features for in-hospital mortality prediction. However, there are many other potential uses of respiratory support information, such as endpoint identification, patient sub-phenotype discovery, patient trajectory analytics, or characterization of patient cohorts.[93, 94, 104]

As documentation regarding respiratory support status varies across time and across sites, identification of timestamped documentation that serve as evidence for respiratory support cannot be generalized and must be specified in each study by researchers with appropriate knowledge of local practice patterns contained within the dataset. Therefore, a (meta-)heuristic was developed which provides the structure for developing heuristics to establish episodes for any respiratory support method.

As is typical of studies using EHR data, this study suffers from missingness and inaccuracy of information. For example, we identified patients admitted through the ED with a recent positive COVID-19 test who were transferred to and remained in the ICU for several days, yet had no documentation of respiratory support throughout their entire stay. While heuristics can work with scattered, conflicting, and incomplete documentation, significant missingness will still result in unrealistic scenarios. There is trade-off in setting the MIN_DURATION parameter – if it's too long then temporary/interim respiratory supports will be underrepresented, conversely, if it's too short then the heuristic will allow for unrealistically rapid oscillation among respiratory support methods. Additionally, tracheostomy status was considered orthogonal to the proposed system. NIMV is often used nightly for sleep apnea; thus, researchers utilizing the heuristic must decide whether to ignore those episodes based on their needs. Also, patients can be connected to a device for respiratory support, but not be actively using them (e.g., delivering no or low flow oxygen through a device capable of delivering high flow oxygen), thus researchers must decide which is more important for their work: the occupation of the device or the active use of the device. Respiratory support methods are ever-evolving – helmet NIV and high-flow nasal cannula, for instance, have only recently been used for adult patients, meaning that these terminology systems will also require regular revisiting and updating.[105]

2.5.5. Conclusion

To facilitate generalizable and reproducible research, a terminology system was developed for standardized representation of respiratory support methods. (Meta-)heuristics were also developed to enable extraction of respiratory support episodes from EHR data, and transformation into encounter-spanning set of respiratory support trajectories. To demonstrate the utility of respiratory support information extracted through proposed methods, feature ablation and feature importance analyses were performed via an in-hospital mortality prediction study for COVID-19-tested, ED-admit, adult patients. The addition of features generated from the proposed approach significantly improved model performance. Further, those features were found to be among the most important for models. Finally, the proposed approach, which generated more interpretable and generalizable representations, despite the loss of conceptual and temporal granularity, had commensurate performance to alternate representations of explicit respiratory support information.

2.6. Discussion and Conclusion

The aim of this chapter was to identify, assess the impact of, and propose solutions for data shift resulting in the non-generalizability and transportability of CPMs. While there are numerous potential causes of data shift general to ML modeling, the focus of this chapter was on CPM-specific causes. Literature review revealed two unique types of causes of ML model non-transportability specific to CPMs: 1) bias in EHR data caused by HCP, and 2) insufficient coverage of EHR data by standards. Therefore, two studies were conducted for each of the two types of causes of CPM non-transportability. The first study focused on the deleterious impact of a unique class of CPM features known as HCP features, finding that they harm external generalizability while improving internal validity and temporal generalizability. Thus, for CPM transportability, models heavily reliant on HCP features should be used with caution – either deciding not to use the CPM or planning on re-training or fine-tuning. In the second study, the lack of coverage by standards was identified as a source of semantic discrepancy of CPM features, so a novel standard was developed and proposed to reduce the risk of non-transportability of respiratory-support-based features. The individual contributions of the two studies are: 1) discovery of HCP features as a critical driver of CPM non-generalizability; and 2) a novel classification system and heuristics for respiratory support methods. The contribution to the APT checklist is the finding that features should be mapped to pre-existing, well-validated, and widely-adopted standards when possible. Otherwise, features must be mapped to ad-hoc standards with clear documentation of the coding process to avoid semantic discrepancy of features between sites and minimize the risk of CPM non-transportability.

Chapter 3. Target Disparity

3.1. Introduction

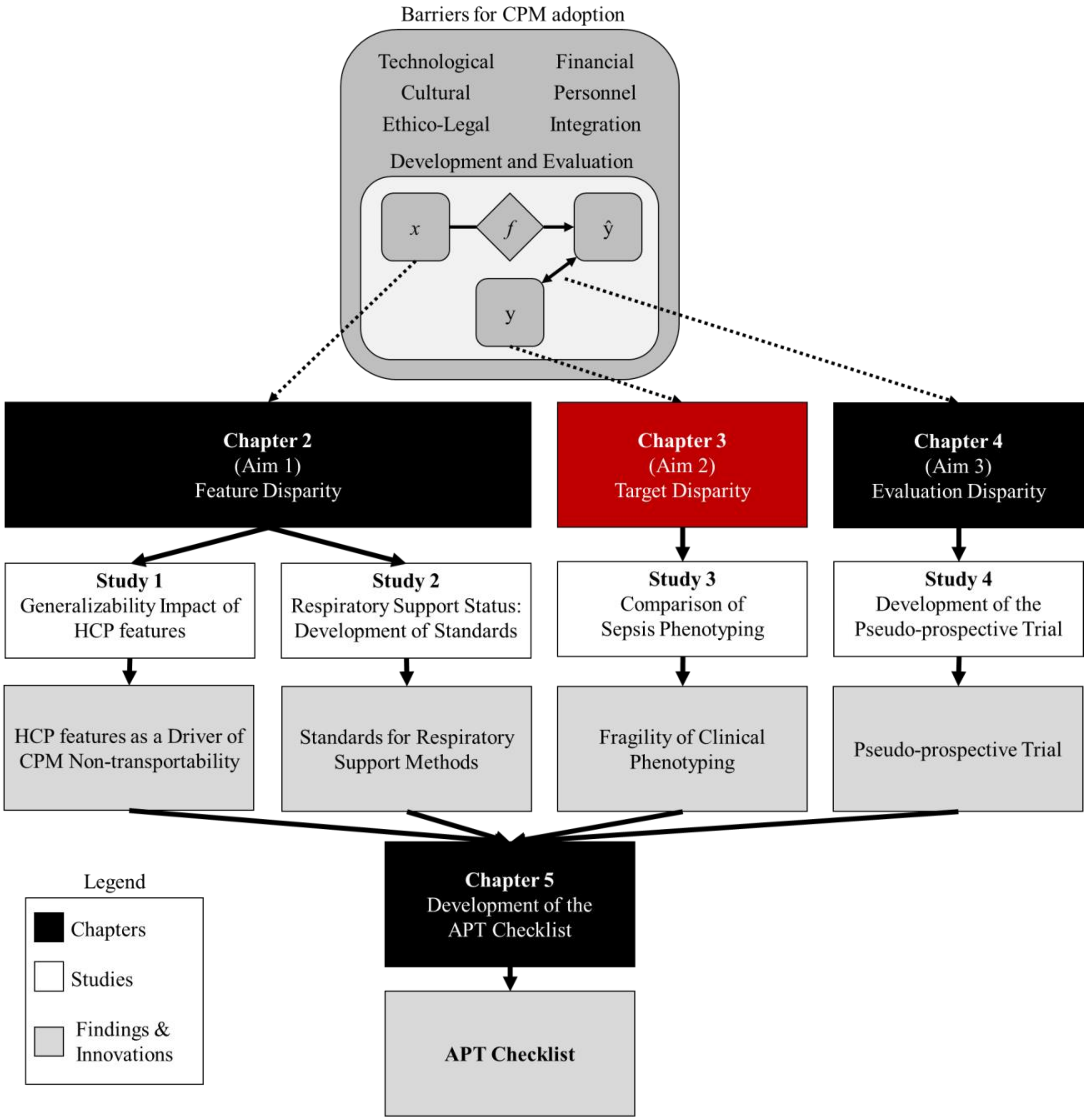

Barriers for CPM adoption
Technological
Financial
Cultural
Personnel
Ethico-Legal
Integration
Development and Evaluation
x
f
ŷ
y
Chapter 2
(Aim 1)
Feature Disparity
Chapter 3
(Aim 2)
Target Disparity
Chapter 4
(Aim 3)
Evaluation Disparity
Study 1
Generalizability Impact of HCP features
Study 2
Respiratory Support Status: Development of Standards
Study 3
Comparison of Sepsis Phenotyping
Study 4
Development of the Pseudo-prospective Trial
HCP features as a Driver of CPM Non-transportability
Standards for Respiratory Support Methods
Fragility of Clinical Phenotyping
Pseudo-prospective Trial
Legend
Chapters
Studies
Findings & Innovations
Chapter 5
Development of the APT Checklist
APT Checklist

Figure 16. Chapter 3 Overview

This chapter addresses specific aim 2, the objective of which is to identify the causes and characterize the impact of heterogeneity in labels required for CPM development and to propose solutions for challenges to transportability of CPMs. To this end, a study was conducted assessing the causes of heterogeneity in the target variable assignment process and its impact on the characteristics of the resultant cohorts. Sepsis was selected as the clinical context because it is a common target for the prediction of CPMs, a syndromic condition with no gold standard test, and because there are many competing yet widely used phenotyping approaches. Through this study, it has been found that there are sources of phenotyping disagreement on both the macro level (disease concept understanding) and the micro level (details of the criteria). Further, these differences manifest in highly heterogeneous populations – all ostensibly septic – with significantly different characteristics and clinical outcomes. So, CPM adopting organizations must acknowledge the fragility of clinical phenotyping used to identify target labels, decide if they agree with the approach used by the CPM development team, and if not, understand the ramifications of the differences in both disease concept and specific criteria. These findings and innovations of the study conducted in this chapter are then used to supplement the development of the APT checklist.

3.2. Overview

Structurally, this chapter will begin by reiterating the motivating specific aims as has already been done in the preceding section, followed by a background section on the following topics – disease classification and clinical phenotyping. Then, the study of this chapter is presented, including its own introduction, background, methods, results, discussion, and conclusion sections. Finally, the chapter concludes by discussing the ramifications of the findings and innovations of the studies on CPM transportability and the APT checklist.

3.3. Background and Significance

3.3.1. ICD Codes

The elephant in the room regarding information about patient conditions or disease diagnosis in EHR data is the reliability of ICD code data – whether the list of codes assigned at discharge is accurate and comprehensive and thus is suitable for secondary research, including CPM development and evaluation. While the study of disease classification has a long and storied history, the healthcare industry in the United States has converged on the ICD system to classify and code diseases.[57] Disease diagnosis information is well-documented – for billing purposes – and adheres to a widely-adopted standard; however, it suffers from the following critical problems. First, ICD codes are documented primarily at discharge and do not contain when within an encounter the patient suffered the onset of a disease. Nevertheless, onset time is critically important for CPMs targeted at the acute care setting. Secondly, ICD codes suffer biases based on their entanglement with hospital billing. For example, researchers observed the "Will-Rogers" phenomenon in pneumonia and sepsis in which severely ill pneumonia patients tended to be classified instead as sepsis (upcoded), resulting in apparent improvement in outcomes for both the pneumonia cohort and sepsis cohort.[63-65] The "gaming" or at least the dual-role – clinical and billing – aspect of diagnosis code assignment results in seemingly paradoxical findings; for example, contrary to common knowledge, obesity is considered protective across numerous comorbidity indices and associated scoring systems likely because mild or even severe obesity is unlikely to be documented especially for critically ill patients with a litany of more clinically significant conditions.[106-110] For these reasons, ICD codes in EHR data cannot be considered sufficient for phenotyping patients. As a result, many researcher networks such as Electronic Medical Records & Genomics (**eMERGE**), Pharmacogenomics Research

Network (**PGRN**), and the Food and Drug Administration's (**FDA**) Mini-Sentinel surveillance initiative have developed phenotypic algorithms that incorporate data available in the EHR beyond ICD codes.[111]

These computational phenotyping algorithms, or from the perspective of CPM development and evaluation, the target variable labeling processes, also suffer from concerns regarding semantic heterogeneity that can result in CPM non-transportability just as how feature semantic discrepancy was shown to result in CPM non-transportability in chapter 2. There are two levels in which organizations can differ regarding the target variable labeling process: 1) the overarching disease concept, and 2) specifics of the phenotyping criteria (**Figure 17**). The following sections provide background information on each of these sources of target disparity.

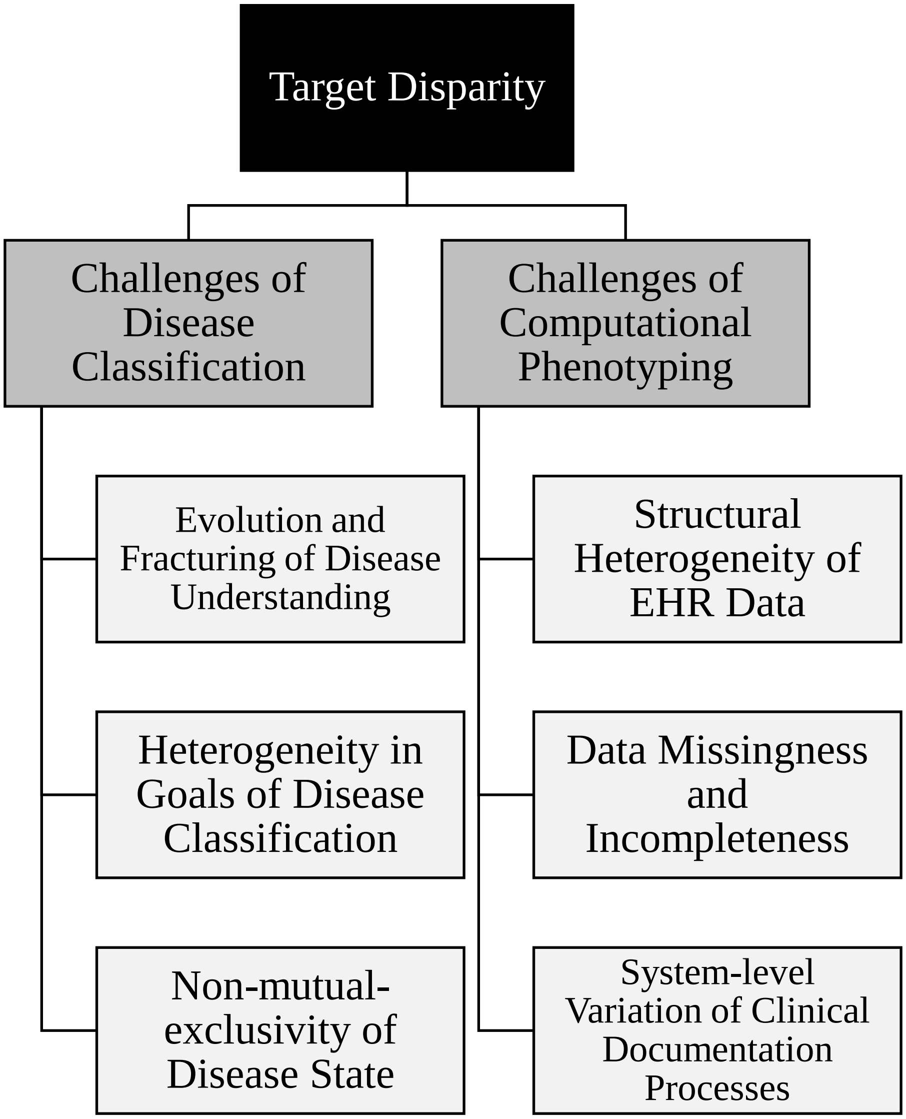

Target Disparity
Challenges of Disease Classification
Challenges of Computational Phenotyping
Evolution and Fracturing of Disease Understanding
Heterogeneity in Goals of Disease Classification
Non-mutual-exclusivity of Disease State
Structural Heterogeneity of EHR Data
Data Missingness and Incompleteness
System-level Variation of Clinical Documentation Processes

Figure 17. Sources of Target Disparity

Conceptual categorization of the challenges to disease classification.

3.3.2. Disease Classification

Target disparity, which limits CPM transportability, can emerge from discrepancies in disease concept understanding based on the challenges of disease classification. While there are other frameworks for conceptualizing challenges to disease classification, we focus on the formulation laid out by Angus et al., who describe the challenges as being three-fold: problems of 1) knowledge, 2) purpose, and 3) statistics.[112] First, due to technological and scientific advances, the conceptual understanding of disease in general and of specific diseases are constantly evolving, resulting in the natural fracturing of disease definitions. Second, there are diverse purposes for disease classification, each of which may benefit most from a definition tailored toward that purpose. Thus, the diversity of purpose contributes to the heterogeneity of definitions. Third is the problem of mutual exclusivity – disease cases often lie between categorizations revealing that clear-cut classifications are but useful fiction. Together, these challenges introduce significant heterogeneity in disease definition and criteria, resulting in target label disparity between organizations, harming CPM transportability. The study conducted in this chapter demonstrates the impact of disease concept heterogeneity on cohort characteristics and clinical outcomes.

3.3.3. Computational Phenotyping

In addition to discrepancies in disease understanding stemming from the challenges of disease classification, seemingly minor differences in phenotyping criteria even among those that agree on the disease concept can lead to disparities in the target variable labels, resulting in CPM non-transportability. Because ICD codes cannot be fully trusted as complete and accurate representations of patient conditions as described in a previous section, and because there are rarely ground-truth phenotypic labels in EHR data, computational phenotyping is performed by applying an algorithmic rule on a constellation of data elements.[111] Even when there is agreement on the overarching disease concept – something that is not guaranteed as described in the previous section – the developers of phenotyping algorithms may naturally specify the criteria differently.[113] In addition, heterogeneity of EHR data in general, as explored in chapter 2, can induce algorithm developers to prioritize certain data elements or methods over others based on the availability, reliability, and likelihood of bias (**Figure 17**).[114, 115] Data-driven, ML-based phenotyping efforts are being explored, but has not seen wide adoption.[116] So, in addition to disparities in disease understanding as described in the previous section, disparities in criteria implementation can contribute to target disparity resulting in CPM non-transportability. The study conducted as part of this chapter explores these particular barriers to CPM non-transportability under the clinical context of sepsis. Sepsis is a complex, syndromic condition lacking a gold-standard test, whose definition has evolved significantly in recent years, and has received a lot of attention due to its high prevalence and mortality rate, including being a common target of prediction for CPM studies.[117, 118] In the following section, the sepsis phenotyping study is described in which various sepsis definitions and criteria are compared, as well as a comparison of the corresponding sepsis cohorts, all to better understand the potential non-transportability impact of CPMs.

3.4. Study 3: Comparison of Sepsis Phenotyping

This study examines the sources and impact of phenotyping heterogeneity in sepsis through the comparison of competing sepsis definitions/criteria as well as the cohorts derived from those definitions/criteria.

3.4.1. Introduction

Sepsis, an exaggerated immune response to an infectious process that can lead to life-threatening organ failure, affects more than 1 million patients each year across the world and carries a high risk of death.[118, 119] Although accounting for a small proportion of all hospital admissions, sepsis is the most expensive condition treated in US hospitals, costing nearly $24 billion on an annual basis.[119] In order to facilitate consistent, large-scale research into sepsis pathobiology and outcomes and to augment clinical trial enrollment, a standardized, automated approach to case identification from EHR is needed.[120] Unfortunately, the lack of a "gold standard" diagnostic test for sepsis precludes easy case identification and necessitates identification through surrogate measures, a process that has been iteratively refined over decades.[118, 121, 122]

In 1992, the American College of Chest Physicians and the Society of Critical Care Medicine defined sepsis as Systemic Inflammatory Response Syndrome (**SIRS**) in response to infection.[121] Although easy to implement bedside, these criteria were found to be non-specific and insufficiently sensitive.[123-125] This definition was revisited in 2001, but the fundamental concept of sepsis as a systemic inflammatory response due to infection did not change.[122] In 2015, the Centers for Medicare and Medicaid Services (**CMS**) introduced the SEP-1 quality metric, where severe sepsis was defined as evidence of infection, in conjunction with SIRS and organ dysfunction.[126] In 2016, the Sepsis-3 consensus definition recharacterized sepsis as organ dysfunction caused by a dysregulated response to infection. The following year, Rhee et al. modified Sepsis-3 to become the Adult Sepsis Event (**ASE**) criteria, which was adopted by the Centers for Disease Control and Prevention (**CDC**).[118, 127, 128]

In response to the changing paradigm of sepsis, several efforts have been made to characterize the shift in diagnostic criteria.[112, 129-131] While many focused on the comparative ability of SIRS and quick Sequential Organ Failure Assessment (**qSOFA**) to identify sepsis or predict mortality, others focused on defining the degree of overlap between these criteria.[132-137] Smaller studies investigating the degree of patient overlap each compared different definitions across different healthcare settings, making comparative analysis difficult. Few studies directly compared more than two criteria using EHR data and often focused on a particular subpopulation or did not evaluate the agreement among sepsis cohorts.[138, 139]

In order to better understand the population-level characteristics of the shifting sepsis criteria, this study simultaneously implemented established sepsis criteria (Sepsis-1, Sepsis-3, CMS SEP-1, and CDC ASE) on a large inpatient population to determine the impact of diagnostic criteria on population characteristics and quantify the occurrence of mortality on different sepsis phenotypes that meet different diagnostic criteria.

3.4.2. Methods

3.4.2.1. Study Design, Data Sources, and Population

This retrospective analysis was conducted using EHR and administrative claims data from Barnes-Jewish Hospital / Washington University School of Medicine in St. Louis, a large, academic, tertiary-care referral center. Eligible patients were at least 18 years of age and admitted to the hospital as inpatients or observation status between 1/1/2012 and 6/1/2018. Patients were excluded if they were admitted to the Psychiatry or Obstetrics services, due to highly variable rates of physiologic data collection. Encounters were excluded if there were no billing code, vital sign, laboratory, service, room, or medication data. Sepsis criteria were evaluated on full patient encounters including the ED, general ward, and ICU setting. Only the first occurrence of sepsis per encounter was included in each analysis. Patients were not eligible to meet sepsis criteria for 72 hours after the conclusion of a surgical procedure to avoid conflation of post-surgical patient status with sepsis. All analysis were performed on a per-encounter level. This project was approved with a waiver of informed consent by the Washington University in St. Louis Institutional Review Board (IRB #201804121).

3.4.2.2. Implementation

Data elements necessary for sepsis cohort identification were mapped from extracted EHR data (**Appendix 24**). Comorbidities were identified using the Elixhauser comorbidity index.[107, 108]

3.4.2.3. Sepsis Definitions

Sepsis criteria were implemented according to consensus definitions with modifications as indicated (**Table 9, Appendix 26, Appendix 27, Appendix 28, and Appendix 29**).[117, 118, 121, 126-128] Sepsis was Present On Admission (**POA**) if the encounter had an admitting diagnosis code of sepsis, severe sepsis, or septic shock (**Appendix 25**).

Table 9. Compared sepsis definitions

Definition	Infection	Anti-infectives	Cultures	Response to Infection	Time constraint	Time zero
Sepsis-1 (Bone, 1992)	Concomitant cultures and anti-infectives	All intravenous antibiotics, antivirals, and antifungals as well as enteral vancomycin and metronidazole.	"Clinical cultures" including bacterial, fungal, viral cultures[a]	Systematic Inflammatory Response Syndrome (SIRS) within a 1-hour window	Cultures followed by anti-infective within 48 hours or anti-infectives followed by cultures within 24 hours. SIRS met between 48 hours before to 24 hours after earlier of either culture or anti-infective	Earlier of either culture collection or anti-infective initiation
CMS SEP-1 (National Hospital Inpatient Quality Measures)	Antibiotics	All intravenous antibiotics (not antifungals or antivirals) as well as enteral vancomycin and metronidazole	N/A	SIRS and organ dysfunction within 6-hour window	All criteria met within a 6-hour window	Latest of met criteria
Sepsis-3 (Seymour, 2016)	Concomitant cultures and anti-infectives	All oral and IV anti-infectives except one-time or perioperative anti-infectives	All bacterial, fungal, viral and parasitic cultures as well as C. diff assays	Sequential Organ Failure Assessment (SOFA) in the critical care setting, and quick SOFA (qSOFA) elsewhere	Cultures followed by anti-infective within 72 hours or anti-infectives followed by cultures within 24 hours. (q)SOFA met between 48 hours before to 24 hours after earlier of either culture or anti-infective	Earlier of either culture collection or anti-infective initiation
CDC ASE (Rhee, 2017)	Concomitant blood culture and anti-infectives	Intravenous and enteral antibiotics, antifungals, and antivirals.	Blood cultures	Acute organ dysfunction (modified SOFA)	Anti-infective initiation and sign of acute organ dysfunction both within 48 hours of blood culture	Blood culture collection

Underlining indicates definition subcomponents that were modified or improvised due to under-specification or to enable automated execution

For further definition implementation details, see Appendix 26 through 29

[a]Clinical cultures as defined by Rhee et al.[128]

International Classification of Disease (ICD) code identified cases of sepsis were recognized using explicit diagnostic codes for sepsis, severe sepsis, and septic shock (**Appendix 25**).[117]

Sepsis-1 case recognition was based on the 1992 consensus guidelines, requiring both Suspicion Of Infection (**SOI**) and at least 2 positive SIRS measurements.[121] SOI was defined as concomitant antibiotics and cultures (**Appendix 26**). The SIRS criteria were defined as at least two of the following within 1 hour: temperature >38.0 C or <36.0 C; heart rate >90; respiratory rate >20 per minute; white blood cell count >12,000 or <4,000 or >10% bands. Onset time was defined as the earlier time of either culture or antibiotics.

CMS SEP-1 was adapted from the severe sepsis definition in the CMS sepsis core measure guidelines, which included SOI, positive SIRS criteria, and at least 1 sign of organ dysfunction – all met within a 6-hour window.[126] Organ dysfunction included shock, acute respiratory failure, acute kidney or hepatic injury, thrombocytopenia, coagulopathy and an elevated lactate (**Appendix 27**). In order to enable automated surveillance, SOI was determined based on antibiotic administration (**Appendix 27**). Because CMS SEP-1 was intended to surveil bacterial sepsis, antivirals, antifungals, and antiparasitics were not considered for SOI.[140] Time of onset, per CMS guidelines, was defined as the time when last of the criteria were met.

Sepsis-3 was defined according to the Sepsis-3 consensus criteria as SOI with either a qSOFA score ≥ 2 in the non-ICU setting or a Sequential Organ Failure Assessment (**SOFA**) score ≥ 2 in the ICU (**Appendix 28**).[118, 127] In accordance with the consensus definition, SOI was defined as concomitant cultures and antibiotics, and the time of infection was defined as the earlier of either cultures or antibiotics.[118, 127] Time of onset was defined as time of infection.

CDC ASE criteria were implemented according to published criteria, with SOI defined as blood culture procurement with concomitant initiation of an antimicrobial regimen of at least 4 qualifying days (**QADs**) and acute organ dysfunction identified through a modified SOFA score (**Appendix 29**)[11.] Organ dysfunction was defined as ≥1 of the following: the requirement for vasopressors or mechanical ventilation for ≥24 hours, acute kidney or hepatic injury, thrombocytopenia, and an elevated lactate level. Time of onset was defined as time of blood culture procurement.

3.4.2.4. Statistical Analysis

Missing data imputation was not performed because 1) data missingness is most pertinent among laboratory values (**Appendix 14**) and missingness in this context is presumably due to clinical decision-making rather than data missing at random; and 2) missing data imputation for heterogeneous, longitudinal data introduces implementation variability that limits interpretability and reproducibility.[128]

Variables were summarized using frequencies and proportions for categorical data or medians and InterQuartile Ranges (**IQR**) for continuous data. Group-wise comparisons were performed using the Kruskal-Wallis, χ^2, and Mann-Whitney U tests where appropriate. A p-value <.01 was considered statistically significant. Analysis and figure generation were performed with Python version 3.7.1 (https://www.python.org/) using Jupyter Notebook (Project Jupyter, https://jupyter.org).

3.4.3. Results

3.4.3.1. Study Population

Over the 89-month study period, 343,977 unique inpatient encounters were recorded among which 286,759 met inclusion criteria (**Figure 18**). The median age for the entire population was 59.2 years (47.0 - 69.5) and 47.6% were female (**Table 10**). Comorbidities in the general population were common (Appendix 15), including diabetes (27.1%), congestive heart failure (20.1%), chronic pulmonary disease (24.4%), chronic kidney disease (19.9%), and cancer (17.4%). The size of sepsis cohorts identified by the different criteria varied significantly (*p<.01*): ranging from 12,494 (4.4%) for CDC ASE to 32,369 (11.3%) for Sepsis-1; as did rate of POA sepsis (*p<.01*) from 11.9% for Sepsis-1 to 28.6% for ICD code; and Elixhauser comorbidity score (*p<.01*) from 16 [6 – 26] for Sepsis-1 to 20 [11-30] for CSM SEP-1.

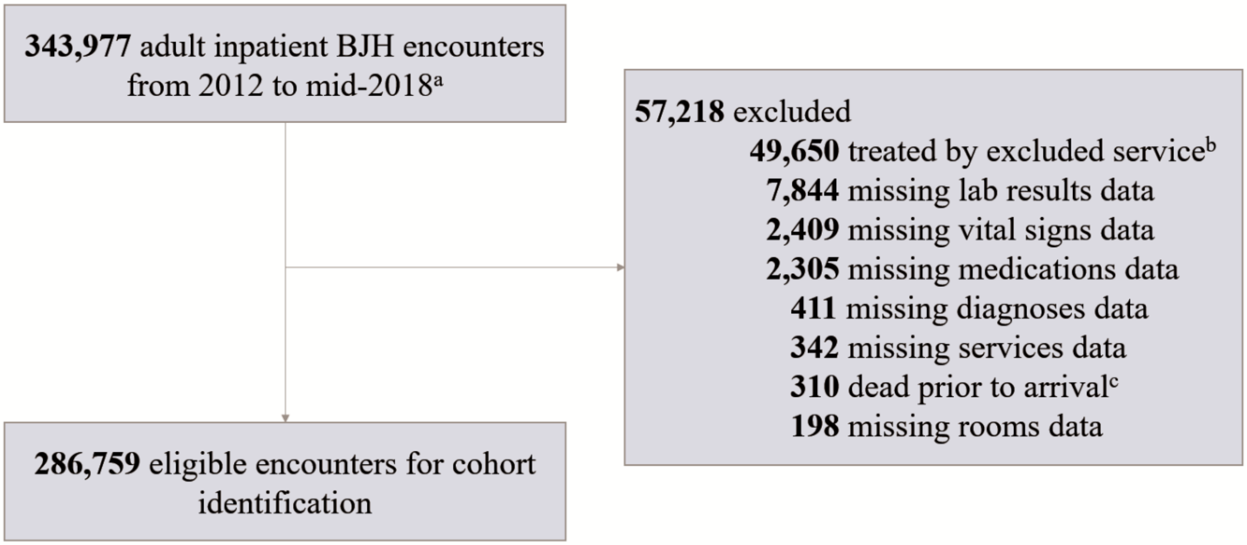

343,977 adult inpatient BJH encounters from 2012 to mid-2018[a]
57,218 excluded
49,650 treated by excluded service[b]
7,844 missing lab results data
2,409 missing vital signs data
2,305 missing medications data
411 missing diagnoses data
342 missing services data
310 dead prior to arrival[c]
198 missing rooms data
286,759 eligible encounters for cohort identification

Figure 18. Patient Enrollment

[a] A single encounter may meet multiple exclusion criteria.
[b] Excluded services included: Obstetrics, Gynecology, Psychiatry, Mental Health, Skilled Nursing, Hospice, Nursery, and Pediatrics.
[c] Date of death preceded recorded admission date.
[d] A method of categorizing medical comorbidities based on billing diagnosis codes. Comorbidity score ranges from -32 to 99.

Table 10. Baseline Population Characteristics

Variable[a]	**Total**	**ICD code**	**Sepsis-1**	**CMS SEP-1**	**Sepsis-3**	**CDC ASE**
Number of encounters (%)	286,759 (100.0%)	20,670 (7.2%)	32,369 (11.3%)	13,869 (4.8%)	21,550 (7.5%)	12,494 (4.4%)
Age, median (IQR), years	59.2 (47.0 - 69.5)	60.8 (49.7 - 70.1)	59.7 (47.7 - 69.4)	60.4 (49.1 - 69.8)	61.8 (51.4 - 71.4)	60.9 (50.6 - 69.5)
Sex (female), No. (%)	136,503 (47.6%)	9,091 (44.0%)	15,048 (46.5%)	6,181 (44.6%)	9,863 (45.8%)	5,513 (44.1%)
Race, No. (%)						
White	187,141 (65.3%)	13,409 (64.9%)	21,310 (65.8%)	9,323 (67.2%)	14,467 (67.1%)	8,398 (67.2%)
Black	82,829 (28.9%)	5,851 (28.3%)	8,902 (27.5%)	3,590 (25.9%)	5,573 (25.9%)	3,244 (26.0%)
Asian[b]	1,867 (0.7%)	149 (0.7%)	227 (0.7%)	84 (0.6%)	146 (0.7%)	94 (0.8%)
Other[b,c]	14,922 (5.2%)	1,261 (6.1%)	1,930 (6.0%)	872 (6.3%)	1,364 (6.3%)	758 (6.1%)
BMI, median (IQR)	27.7 (23.5 - 33.0)	27.4 (23.1 - 33.0)	27.2 (22.9 - 32.8)	27.6 (23.3 - 33.0)	27.5 (23.0 - 33.2)	28.0 (23.5 - 33.7)
Sepsis present on admission, No. (%)	6,128 (2.1%)	5,913 (28.6%)	3,861 (11.9%)	2,141 (15.4%)	2,819 (13.1%)	2,247 (18.0%)
Elixhauser comorbidity score, median (IQR)[d]	7 (0 - 16)	18 (9 - 28)	16 (6 - 26)	20 (11 - 30)	19 (10 - 29)	20 (10 - 29)

Abbreviations: ICD, International Classification of Disease; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event; IQR, interquartile range; BMI, body mass index.

[a] Comparison across all four definition-based cohorts was performed using Kruskal-Wallis one-way analysis of variance test for continuous variables, and χ^2 for categorical variables. All comparisons were statistically significant (*p*<.01) except where denoted by superscript b.
[b] Race: Asian, $\chi^2(3) = 2.25$, $p \geq .01$; Race: Other, $\chi^2(3) = 3.75$, $p \geq .01$.
[c] Other race includes Pacific Islander, Hispanic, Alaska Native, Unknown, Other, and more than one race.
[d] Elixhauser comorbidity score was calculated based on formula from Moore et al.[108]

3.4.3.2. Overlap between sepsis criteria

The level of agreement between sepsis criteria was low for all pairs except between Sepsis-1 and Sepsis-3 where the level of agreement was moderate (κ = 0.533, 95% CI, 0.530 - 0.536;

Appendix 16). Of the 32,369 Sepsis-1 cases (13.6% in-hospital mortality rate) and 21,550 Sepsis-3 cases (18.8%), only 15,508 (47.9% of Sepsis-1, 72.0% of Sepsis-3) met both criteria and had an increased in-hospital mortality rate of 21.5% (**Figure 19**). Only 4,370 encounters met criteria for all definitions and had an in-hospital mortality rate of 37.0%. Overall, in-hospital mortality was higher for populations that met more than 1 definition and increased by an average of 5.4% per additional criteria met ($r^2 = 0.740$; *p<.01*; **Appendix 17**).

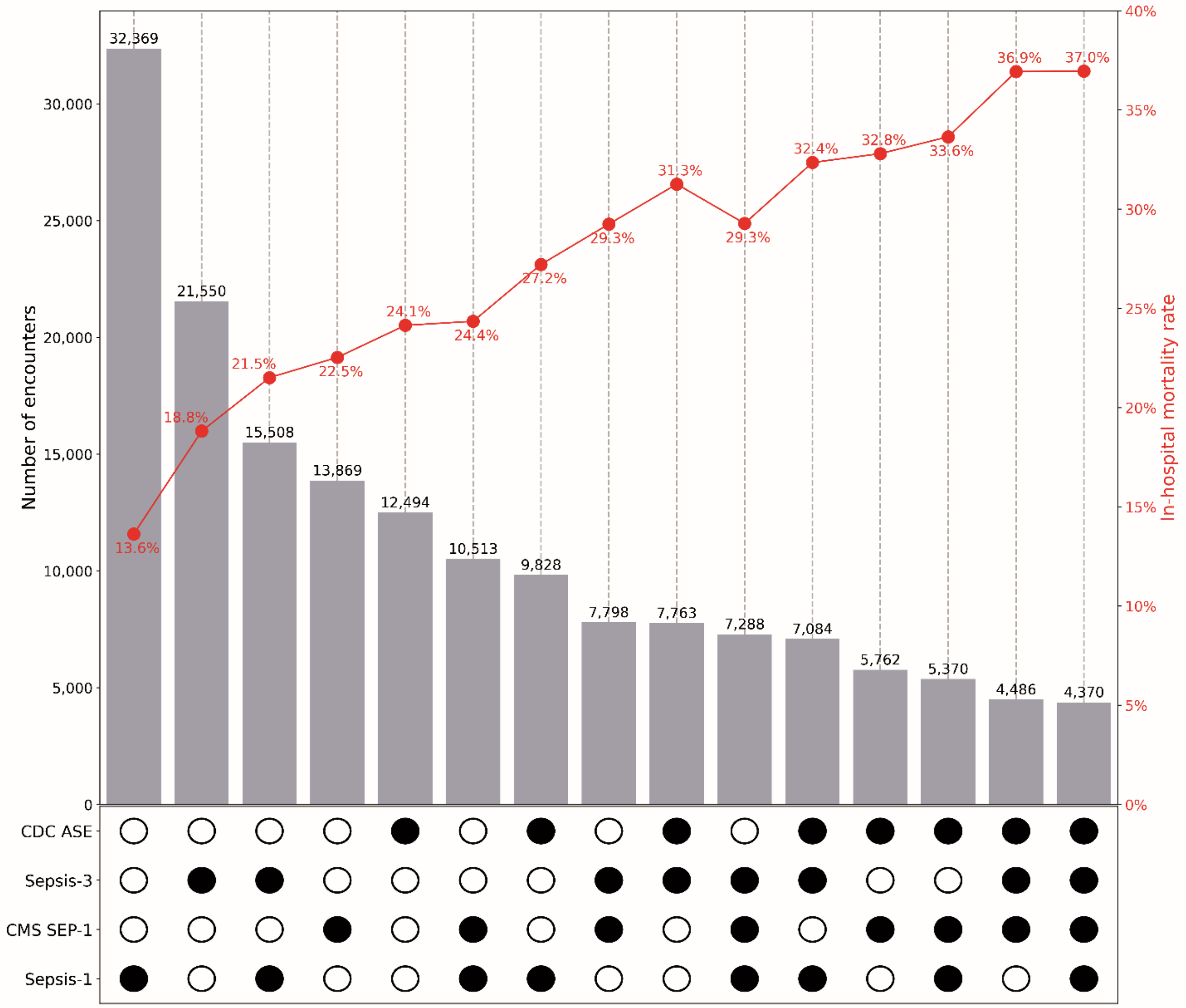


Figure 19. Patient Overlap between Different Sepsis Definitions with Associated Mortality

Filled in circles indicate corresponding cohort(s) represented in the bar above. The bar plot demonstrates the number of encounters meeting all depicted criteria (intersection). The red line plot indicates the in-hospital mortality rate per respective cohort.

Abbreviations: CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event.

3.4.3.3. Clinical differences between sepsis criteria

The time-to-onset (hours) varied significantly across the sepsis definitions ($p< .01$): from 2.9 for Sepsis-1 to 7.6 for CMS SEP-1. The distribution of sepsis onset location also varied significantly between definitions ($p<.01$, **Table 11**). Sepsis onset occurred in the ED for only 1.7% for the

CMS SEP-1 cohort, a significantly lower rate compared to the other definitions (23.5% – 28.1%). The most common onset location for Sepsis-1 was the general ward (45.9%) whereas the ICU was the most common location for Sepsis-3 (42.5%) and CDC ASE (38.8%). At sepsis onset, measures of illness severity both varied significantly (*p<.01*, **Table 11**); APACHE II score varied from 14 [10-17] for Sepsis-1 to 16 [12-20] for CMS SEP-1 and median SOFA score varied from 3 [1-5] for Sepsis-1 to 4 [3-6] for CMS SEP-1 and CDC ASE.

Hospital length of stay (days) varied significantly (*p<.01*, **Table 11**) from 8.3 for Sepsis-1 to 11.3 for CDC ASE; as did the presence of a sepsis discharge ICD code rate (*p<.01*) from 34.7% for Sepsis-1 to 54.5% for CDC ASE; severe sepsis discharge ICD code rate (*p<.01*) from 23.2% for Sepsis-1 to 42.1% for CDC ASE; septic shock discharge ICD code rate (*p<.01*) from 16.7% for Sepsis-1 to 34.6% for CDC ASE; and in-hospital mortality rate (*p<.01*) from 13.6% for Sepsis-1 to 24.1% for CDC ASE.

Table 11. Onset and Outcome-related Measures by Sepsis Definition

Variable[a]	**Total**	**ICD code**	**Sepsis-1**	**CMS SEP-1**	**Sepsis-3**	**CDC ASE**
Length of stay, median (IQR), d	3.9 (2.2 – 6.9)	9.6 (5.0 – 19.7)	8.3 (4.7 – 16.2)	9.8 (5.1 – 19.9)	9.3 (5.1 – 17.7)	11.3 (6.1 – 20.1)
Time to onset, median (IQR), h	-	-	2.9 (1.0 – 16.0)	7.6 (3.4 – 26.3)	4.1 (1.1 – 33.4)	4.6 (1.3 – 31.1)
Location at onset, No. (%)						
ER	-	-	9,099 (28.1%)	235 (1.7%)	5,056 (23.5%)	3,095 (24.8%)
General	-	-	14,843 (45.9%)	5,515 (39.8%)	7,084 (32.9%)	4,362 (34.9%)
ICU	-	-	7,605 (23.5%)	7,932 (57.2%)	9,165 (42.5%)	4,851 (38.8%)
Unknown	-	-	822 (2.5%)	187 (1.3%)	245 (1.1%)	186 (1.5%)
APACHE II score at onset, median (IQR)[b]	-	-	14 (10 – 17)	16 (12– 20)	15 (11– 18)	16 (12– 19)
SOFA score at onset,	-	-	3	4	3	4

median (IQR)[b]			(1– 5)	(3– 6)	(2– 5)	(3– 6)
Time from onset to abx (hours), median (IQR)[c]	-	-	1.3 (0.0 – 7.4)	0.0 (0.0 – 0.0)	1.4 (0.0 – 8.0)	0.1 (0.0 – 4.8)
ICU transfer in the 72h following onset among non-ICU onset patients, No. (%)[d]	-	-	7,860 (31.7%)	1,132 (19.1%)	6,229 (50.3%)	3,760 (49.2%)
Mechanical ventilation initiation in the 72h following onset, No. (%)	-	-	5,016 (15.5%)	1,765 (12.7%)	4,493 (20.8%)	3,596 (28.8%)
Vasopressor initiation in the 72h following onset, No. (%)	-	-	5,630 (17.4%)	2,450 (17.7%)	5,731 (26.6%)	4,187 (33.5%)
Sepsis discharge ICD code, No. (%)	20,670 (7.2%)	20,670 (100.0%)	12,117 (37.4%)	6,725 (48.5%)	8,648 (40.1%)	6,812 (54.5%)
Severe sepsis discharge ICD code, No. (%)	11,273 (3.9%)	11,273 (54.5%)	7,502 (23.2%)	4,894 (35.3%)	6,535 (30.3%)	5,259 (42.1%)
Septic shock discharge ICD code, No. (%)	7,631 (2.7%)	7,631 (36.9%)	5,420 (16.7%)	3,824 (27.6%)	5,132 (23.8%)	4,318 (34.6%)
In-hospital mortality, No. (%)	8,839 (3.1%)	4,209 (20.4%)	4,413 (13.6%)	3,125 (22.5%)	4,055 (18.8%)	3,017 (24.1%)
Discharge disposition, No. (%)						
Discharge to home	232,216 (81.0%)	10,397 (50.3%)	19,314 (59.7%)	6,876 (49.6%)	10,704 (49.7%)	5,999 (48.0%)
Discharge/transfer to nonacute care facility[e]	39,566 (13.8%)	5,063 (24.5%)	7,239 (22.4%)	3,209 (23.1%)	5,781 (26.8%)	2,923 (23.4%)
Discharge/transfer to acute care hospital[e]	1,988 (0.7%)	292 (1.4%)	375 (1.2%)	164 (1.2%)	256 (1.2%)	150 (1.2%)
Discharge to hospice facility[e]	3,313 (1.2%)	665 (3.2%)	942 (2.9%)	454 (3.3%)	686 (3.2%)	369 (3.0%)
Miscellaneous/other[e]	837 (0.3%)	44 (0.2%)	86 (0.3%)	41 (0.3%)	68 (0.3%)	36 (0.3%)
30-day readmission, No. (%)	49,535 (17.3%)	4,293 (20.8%)	7,100 (21.9%)	2,961 (21.3%)	4,254 (19.7%)	2,753 (22.0%)

Abbreviations: IQR, interquartile range; ICD, International Classification of Disease; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event; APACHE, Acute Physiology and Chronic Health Evaluation; SOFA, Sequential Organ Failure Assessment; abx, antibiotics; ICU, intensive care unit; MV, mechanical ventilation.

[a] Comparison across all four definition-based cohorts was performed using Kruskal-Wallis one-way analysis of variance test for continuous variables, and χ^2 for categorical variables. All comparisons were statistically significant (Length of stay, Kruskal-Wallis H[3] = 740.99, *p*<.01; Time to onset, Kruskal-Wallis H[3] = 2512.74, *p*<.01; APACHE II score at onset, Kruskal-Wallis H[3] = 991.31, *p*<.01; SOFA score at onset, Kruskal-Wallis H[3] = 2339.87, p<.01; Time from onset to antibiotics, Kruskal-Wallis H[3] >10000, *p*<.01; ICU transfer in the 72h following onset, χ^2[3] = 2524.51, *p*<.01; Mechanical ventilation initiation in the 72h following onset, χ^2[3] = 1453.85, *p*<.01; Vasopressor initiation in the 72h following onset, χ^2[3] = 1750.23, *p*<.01; Sepsis discharge ICD code, χ^2[3] = 1328.50, *p*<.01; In-hospital mortality, χ^2[3] = 930.69, *p*<.01; Discharge to home, χ^2[3] = 858.57,

p<.01; 30-day readmission, $\chi^2[3] = 43.08$, p<.01) except where denoted by superscript e.
[b] Scores were calculated only for those with sufficient data which was defined as at least one measurement of each of the following variables in the 24h preceding sepsis onset: heart rate, systolic blood pressure, temperature, respiratory rate, oxygen saturation, white blood cell count, and creatinine.
[c] If patient had already received antibiotics by the time of sepsis onset, then time from onset to antibiotics was set to 0h. Because antibiotics are a component of all compared definitions, time from onset to antibiotics is sensitive to the definition of onset, especially its temporal relationship with time of antibiotics. Notably, time to onset is 0h by definition for CMS SEP-1.
[d] Encounters where patient was already in ICU at time of onset were excluded.
[e] Discharge disposition: hospice, $\chi^2[3] = 6.16$, $p \geq 0.01$; discharge disposition: acute care hospital, $\chi^2[3] = 0.18$, $p \geq 0.01$; discharge disposition: miscellaneous, $\chi^2[3] = 1.16$, $p \geq 0.01$.

3.4.4. Discussion

In this retrospective analysis, commonly used sepsis criteria were adapted for automated prospective case recognition and applied to patients admitted to the general ward. The resultant populations showed significant heterogeneity in size as well as patient characteristics, with the SIRS criteria representing the largest and least critically ill population with the CDC ASE criteria identifying the smallest cohort with the highest mortality. Moreover, there was little agreement between established sepsis criteria, however, when patients met multiple criteria for sepsis, mortality increased, rising by 5% for additional criteria met, on average.

This study reinforces the growing body of literature that suggests no singular definition of sepsis encapsulates the entire spectrum of disease severity and highlights the population heterogeneity seen with varying sepsis criteria.[132, 133, 135, 136, 138, 139] This study adds to existing literature by incorporating one of the largest sets of sepsis criteria and applies them across a variety of care settings in one of the largest inpatient cohorts. In addition, this analysis uniquely quantifies the degree of agreement between criteria and provides for the first time a quantification of the mortality seen by different phenotypes of sepsis that meet different overlapping criteria.

Surprisingly, low concordance was found between the two interpretations of the Sepsis-3 definition (Sepsis-3 and CDC ASE, $\kappa = 0.424$), where 64.0% of the Sepsis-3 cohort did not meet CDC ASE criteria and 37.8% vice versa. The differences in clinical criteria that give rise to this heterogeneity warrants further discussion. For instance, the SOI criteria for the CDC ASE definition requires the administration of at least 1 dose of an intravenous antibiotic and blood culture procurement, whereas the Sepsis-3 criteria permits both parenteral and enteral antibiotics as well as a much wider array of cultures. Further, the CDC ASE criteria includes shock and the need for mechanical ventilation, whereas Sepsis-3 necessitates at least a 2-point change in SOFA score. The low concordance between CMS SEP-1 and other definitions is unsurprising given the differences in underlying sepsis concepts, however, even within-criteria variability has been observed with CMS SEP-1 ($\kappa = 0.40$).[126, 141] Cumulatively, these differences manifest in cohorts with significantly different characteristics and clinical outcomes, which has far reaching implications on sepsis surveillance, phenotyping and outcomes research.[112, 139] The path forward may be to recognize that competing sepsis criteria yield distinct phenotypes, and that the choice of criteria should depend mostly on the intended use case.[112]

The compared definitions varied in the ease with which they could be automated and the degree to which modifications were necessary. Sepsis-3 and CDC ASE were relatively easy to automate given the level of specificity provided in the original manuscripts. However, some differences in specificity remained – for example, CDC ASE explicitly lists the names of qualifying anti-infectives for SOI whereas Sepsis-3 provides broad categories (oral or parenteral). Without a consistent and widely adopted classification system for anti-infectives and cultures, heterogeneity in interpretation is unavoidable. Sepsis-1 was the least detailed and required the most interpretation. For instance, the time window within which at least 2 SIRS criteria must be met was not specified. Consistent with the work by Churpek et al, we decided that SIRS elements should occur relatively simultaneously, thus settled on a 1-hour window.[124] CMS SEP-1 was highly detailed, but designed for manual chart review, thus significant changes had to be made to enable automated execution which primarily involved simplifying the criteria to use discrete data. Further, due to the heterogeneity in sepsis diagnosis code application, the preliminary filtering step using ICD codes was omitted.[128] Unfortunately, these changes manifested in the results being incongruous with prior reports. The CMS SEP-1 cohort had a mortality rate of 23% which is higher than previous reports of 16%.[117, 139, 142, 143] Notably, the proportion of sepsis onset in the ED for CMS SEP-1 was only 1.7% compared to the 76.7% found in prior literature, which likely reflects the changes made to the suspicion of infection component of CMS SEP-1 and its impact on time-zero.[142] Ascertainment of time zero for CMS SEP-1, however, is prone to disagreement, in some cases up to a 94.75-hour difference.[144] In order to minimize variability in interpretation, sepsis criteria should be highly specified and algorithmically executable using EHR data.

Identification of sepsis onset time, or time zero, was often under-defined in many sepsis definitions, and thus was inferred based on supporting literature. Unlike other definitions, time zero was explicitly defined under CMS SEP-1, which is the time when all requisite criteria have been met.[126] For CDC ASE, because the definition was anchored around blood cultures, we considered time of sepsis onset as time of blood culture procurement.[128] In Sepsis-1 and Sepsis-3, time of onset was defined as earlier of either culture collection or anti-infective initiation as it represents the earliest time point in the clinical trajectory where sepsis was considered. Because CMS SEP-1 waits until the last time point when all criteria are met, its cohort had a longer time-to-onset and a different distribution of onset location. Differences in cohort characteristics at onset are likely driven by differences in time zero definitions in addition to differences in the cohorts themselves.

The advent of Sepsis-3 has shifted the spectrum of illness severity such that sepsis now refers to patients with end-organ dysfunction, which was previously severe sepsis. CMS SEP-1 operates under this paradigm and surveils for severe sepsis. However, given the aforementioned shift in severity and the erasure of the severe sepsis concept in the modern paradigm, CMS SEP-1 was compared alongside other established sepsis criteria. The variance in implied severity among sepsis definitions likely contribute to the heterogeneity of cohorts identified.

Given the nature of the data collection, processing, and analysis, this study has necessary limitations. First, these results arise from a single, large, tertiary academic medical center, which may preclude generalizability. Second, this study only included adult patients and excluded those admitted to the inpatient psychiatry, obstetrics/gynecology and post-surgical services. More work is necessary in validating the sepsis criteria in these cohorts. Third, though great care was placed in following all established criteria, many definitions required adaptions to enable automated case identification based on EHR data, which could alter definition performance. Fourth, as mental status information was missing from this data set, normal mentation was assumed for all patients, similar to other Sepsis-3 evaluation approaches.[118]

3.4.5. Conclusion

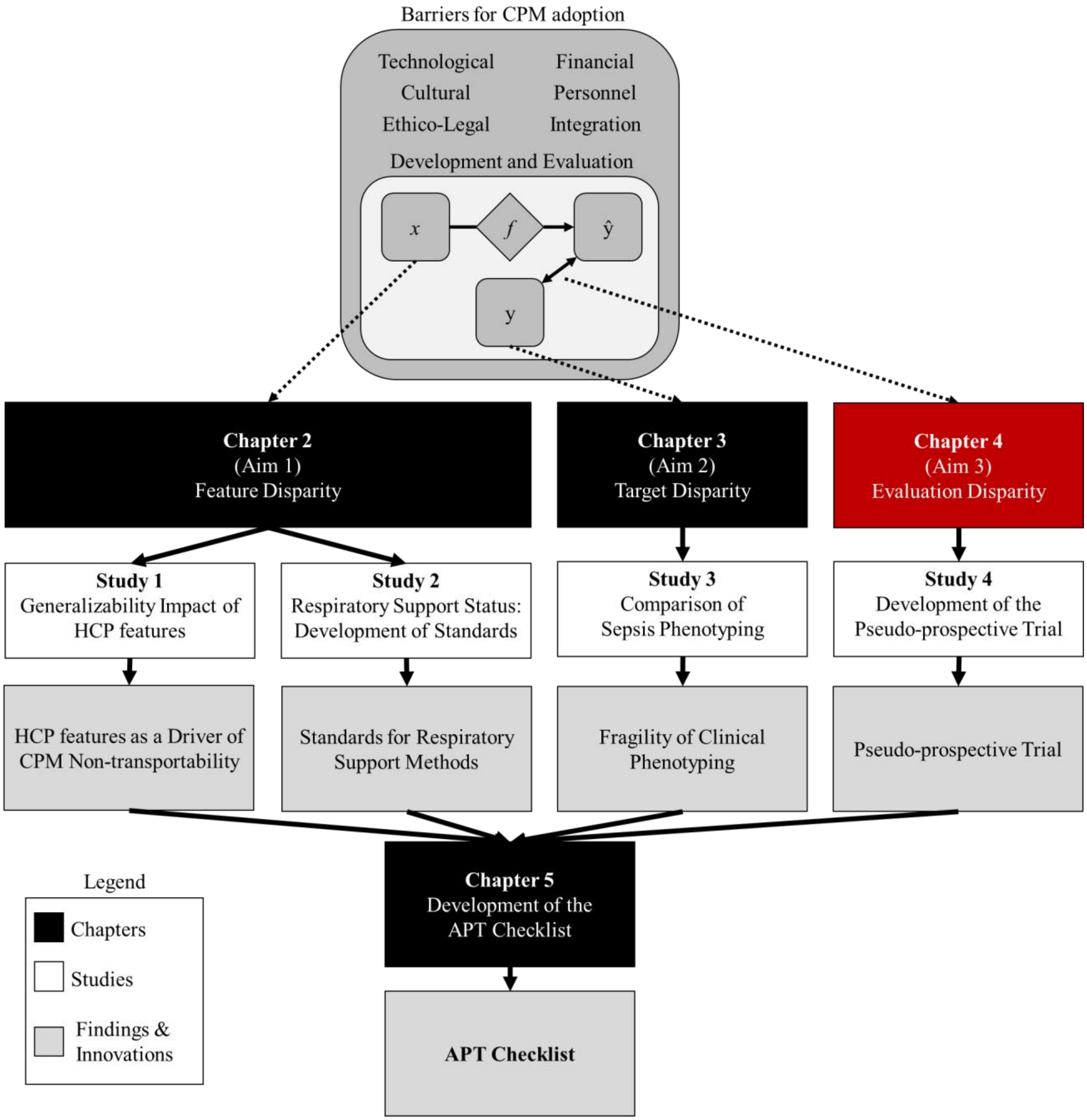
This retrospective analysis of 286,759 encounters from a large, tertiary-referral center revealed that sepsis cohorts based on commonly used sepsis definitions have significantly different characteristics and clinical outcomes and have low concordance with one another. To reduce heterogeneity and improve reproducibility in clinical practice and outcomes-related research, a standardized and automated approach to case identification is needed.

3.5. Discussion and Conclusion

The aim of this chapter was to identify, assess the impact of, and propose solutions for heterogeneity of target variable labeling process caused by variance in clinical phenotyping, resulting in non-generalizability and non-transportability of CPMs. There are prediction targets that are mostly unambiguous – in-hospital mortality for example – but more often, prediction targets are complex clinical phenotypes which was the focus of this chapter. Because of the untrustworthiness of explicit phenotype documentation in EHR data, clinical phenotyping is performed using a constellation of data elements through rule-based, algorithmic criteria. Literature review revealed two categories of causes behind clinical phenotyping heterogeneity: 1) discrepancies in disease understanding, and 2) discrepancies in implementation details of phenotyping criteria. A study was conducted to identify and characterize competing disease definitions and criteria, as well as comparing the various cohorts derived from those competing definitions/criteria regarding their cohort characteristics and clinical outcomes. Sepsis was chosen as it is a complex, syndromic condition with no gold-standard whose definition has evolved significantly multiple times in recent years. It was found that conceptual differences exist even among widely used definitions – e.g., CDC operates under the Sepsis-3 paradigm whereas CMS and ISDA have not yet decided to endorse the new definition and remain under the Sepsis-2 paradigm.[131] Also, there are notably differences between criteria based on even the same definitions – e.g., CDC ASE (Rhee) and Sepsis-3 (Seymour) differ on what are considered acceptable cultures and anti-infectives. These differences result in significant differences in cohorts – not only in terms of cohort size and characteristics but also in terms of clinical outcomes such as in-hospital mortality rates. Given these findings, and given that a CPM trained on one target is unlikely to perform well for a different target, CPM transportability is dependent on the parity between the developer's target labeling process and the adopting organization's approach to clinical phenotyping. Thus, organizations seeking to adopt a CPM must assess the parity of the target variable phenotyping process to identify discrepancies in disease understanding, particulars of the criteria, the cause of the discrepancies, and the impact of the discrepancies on the disease cohort.

Chapter 4. Evaluation Disparity

4.1. Introduction



Barriers for CPM adoption
Technological
Financial
Cultural
Personnel
Ethico-Legal
Integration
Development and Evaluation
x
f
ŷ
y
Chapter 2
(Aim 1)
Feature Disparity
Chapter 3
(Aim 2)
Target Disparity
Chapter 4
(Aim 3)
Evaluation Disparity
Study 1
Generalizability Impact of HCP features
Study 2
Respiratory Support Status: Development of Standards
Study 3
Comparison of Sepsis Phenotyping
Study 4
Development of the Pseudo-prospective Trial
HCP features as a Driver of CPM Non-transportability
Standards for Respiratory Support Methods
Fragility of Clinical Phenotyping
Pseudo-prospective Trial
Chapter 5
Development of the APT Checklist
APT Checklist
Legend
Chapters
Studies
Findings & Innovations

Figure 20. Chapter 4 Overview

This chapter addresses specific aim 3, the objective of which is to characterize and provide solutions for heterogeneity in the framing of CPM evaluation approaches by bridging the gap between CPM evaluation design and expected implemented behavior of CPM-based CDS. Differences in how the implementing organizations think a CPM ought to be and will be evaluated compared to how the CPM was evaluated during the development process – discrepancies in CPM evaluation designs – can result in significant differences in performance, thereby complicating the evaluation of transportability. As such, a study was conducted to bridge the gap between CPM evaluation designs during the development process, and the expected behavior of the CPM when implemented into an external site. This was accomplished through the development of a pseudo-prospective trial concept in which implementation factors such as alert snoozing behavior, when possible, are integrated into the CPM evaluation process. This novel CPM evaluation framework was demonstrated using sepsis prediction in the general ward as the clinical context, and the integration of numerous implementation factors into the evaluation of CPM through the pseudo-prospective trial was found to significantly extend and enrich CPM performance understanding. These findings and innovations of the study conducted in this chapter are then used to supplement the development of the APT checklist.

4.2. Overview

Structurally, this chapter will begin by reiterating the motivating specific aims as has already been done in the preceding section, followed by a background section on the following topics – heterogeneity in CPM evaluation design and simulation-based methods for model evaluation. Then, the study of this chapter is presented, including its own introduction, background, methods, results, discussion, and conclusion sections. Finally, the chapter concludes by discussing the ramifications of the findings and innovations of the studies on CPM transportability and the APT checklist.

4.3. Background and Significance

4.3.1. Heterogeneity in CPM Evaluation Design

Assessment of CPM transportability involves the comparison of model performances – between those on the development set and on the unforeseen external set.[145] There are numerous metrics that can be compared, and for the type of models that are of interest in this dissertation – supervised, binary classification models – the machine learning field has converged on a common set of metrics such as AUROC (**Table 12**).[146] The standardization of performance metrics and integration into machine learning software like the popular packages scikit-learn in python or yardstick in R have facilitated model evaluation and performance understanding.[101, 147] However, these performance metrics are heavily influenced by factors beyond the model or the data, introducing heterogeneity in the meaning of performance metrics that are ostensibly the same, thereby limiting assessment of CPM transportability and thus CPM transportability itself.

A critical source of this heterogeneity is the variability in what we is refer to as the evaluation design in this dissertation – also called the *framing* of prediction models by others – which is about how the evaluation process was designed, which can significantly impact the numerical values of performance metrics.[148] In the case of CPMs in the acute care setting, models can be evaluated once-per-patient or at a patient-level, encounter-level, or multiples times per patient encounter.[32] In addition, these models can look forward at multiple different time scales or time horizons for the presence of an outcome event.[18] In one study externally evaluating the Epic sepsis model, the authors designed and conducted multiple evaluations – at an encounter-level (which they called hospitalization-level) as well as every 15 minutes, assessing for the outcome at 4, 8, 12, and 24 hour time horizons, each yielding their own set of performance metrics, with the AUROCs ranging from 0.63 to 0.76.[18] Just the variability in temporality of evaluation design result in significant heterogeneity of CPMs even with the data and model held unchanged.

Table 12. Traditional Evaluation Metrics for Binary Classification Models

Categories	Metrics
Proper Scoring Function	Log Loss or Binary Cross Entropy Brier Score
Single Metrics	Area Under Receiver Operating Characteristic Curve Area Under Precision Recall Curve
Dichotomization-Based Metrics	True Positive Rate or Recall or Sensitivity Precision or Positive Predictive Value False Positive Rate Positive Predictive Value Negative Predictive Value Accuracy F1 Score
Calibration	Calibration Curve or Reliability Diagram Hosmer-Lemeshow Test Cox Slope and Index

Metrics commonly used for the evaluation of binary classification models

In addition to temporality, there are additional factors that result in evaluation design heterogeneity stemming from the lack of consideration of implemented CPM behavior. CPMs, especially supervised binary classification models, when integrated into clinical practice as CDS, are done so often in the form of interruptive alerts.[149] As a result, there are behaviors specific to CDS alerts that influence the real-world clinical impact of CPMs including but not limited to: conditional and dynamic inclusion or exclusion criteria, lag between data collection and documentation, and alert “snoozing” or lockout periods.[150] These factors not only influence the success of CPM-based CDS interventions, but also add further heterogeneity to the evaluation process as model developers can decide whether or not to take any of these multitude of factors into consideration.

Together, the factors pertaining to the temporality of CPMs and to those pertaining to the behavior of CDS alerts introduce significant heterogeneity of evaluation design, resulting in disparity between CPM evaluation design by the model developers and the expected implemented behavior by those adopting the CPM. To put it another way, AUROC as calculated by the CPM developers may not be how the CPM adopting organization would expect the AUROC to be calculated due to the exclusion of the aforementioned factors into the evaluation process, thereby limiting and confusing expected implemented performance and thus transportability.[151]

The study conducted in this chapter provides a framework for better incorporating these factors into the evaluation design so as to bring parity between the CPM development and evaluation process. The proposed framework is a simulation-based approach, the background for which is provided in the following section.

4.3.2. Simulation-based Methods for CPM Evaluation

As described in the previous section, assessment of CPM transportability can be improved by bringing parity between the CPM evaluation design carried out by the development team and the expected behavior of the CPM-based CDS when implemented in the adopting organization. This can be accomplished through the incorporation of temporal and implementation factors into the evaluation design through a simulation framework. In the context of this chapter, simulation-based methods refer to those akin to Discrete-Event Simulation (**DES**) – the goal is which is to monitor and compare complex system dynamics over time under different conditions or policies – which has been used in healthcare for disease progression, staff scheduling, health screening modeling, and health behavior modeling.[152-157]

Numerous research groups have incorporated DES-based visualization of CPM behavior on individual patients, primarily of exemplary cases of model success, into their manuscripts.[150, 158-160] Often, these visualizations track the behavior of the CPM over time (predicted probability or risk of outcome at regular intervals throughout the patient encounter) alongside other factors such as patient location, important covariates, and critical clinical events including the primary outcome of interest. These visualizations contain information about CPM behavior beyond traditional performance metrics such as the smoothness of continuous predictions, relationship with significant clinical events such as surgical or medical interventions. While these visualizations of simulated CPM behavior are informative, they are primarily used as exemplary demonstrations of CPM success, not as extensions of evaluative design.

Others, however, have augmented CPM evaluation through the incorporation of behaviors specific to CPM-based CDS. One study incorporated alarm silencing regimens directly into their evaluative processes such that while the CPM was executed every 5 minutes, once an alert fires, subsequent alerts would be suppressed for 30 minutes afterwards, thus resulting in a unique evaluation level or frequency distinct from typical CPM studies and more closely reflecting real clinical environments.[150] Instead of focusing on the alarm behavior, some have attended to the limited nature of healthcare resources, thereby deriving an entirely new set of performance metrics that are immediately meaningful to stakeholders; specifically, using surgical ED readmission prediction as the clinical context, the authors derived patients seen, readmissions anticipated, expected readmissions prevented, expected readmission cost, provider cost, expected readmission cost savings, and expected net cost savings – metrics that are additive to traditional performance metrics, and are meaningful for both clinicians and administrators.[161] Another study highlighted the tendency of typical model evaluation strategies to over-inflate model performance, and leveraged a simulation prospective validation strategy to compute metrics more accurate to implemented performance, finding all compared sepsis prediction models to have PPVs below 0.035 at a fixed sensitivity of 50%, which is significantly lower than the 0.15 found in literature.[162] Collectively, there has been a constellation of independent efforts in attempting to incorporate expected implementation behavior into evaluation design to bring parity between CPM evaluation conducted by model developers and the expected behavior of the CDS based on the model once implemented into an adopting organization; yet there has been no framework proposed for uniting these fractured efforts. The study conducted in this chapter addresses this specific gap in literature through the proposal of the pseudo-prospective trial framework, which is described in the following section.

4.4. Study 4: Development of the Pseudo-prospective Trial

This study proposes a novel evaluation design framework for CPMs to improve parity between CPM evaluation design and expected implemented behavior of CPM-based CDS by factoring in the behaviors of the CPM-based CDS into the evaluation design.

4.4.1. Introduction

Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection.[118] In 2017, sepsis was responsible for 5.8% of all hospital stays and $38.2 billion in hospital costs.[163] Moreover, sepsis has a high mortality rate, and was found to be implicated in about one in every three inpatient deaths.[164]

Early and effective therapy is critical in the management of septic patients, as prolonged recognition and delayed treatment increases mortality.[165, 166] As a result, there is an abundance of literature focusing on the early detection and prediction of sepsis through traditional or newly developed scoring systems such as Systemic Inflammatory Response Syndrome (**SIRS**) score, National Early Warning Score (**NEWS**), or quick Sequential Organ Failure Assessment (**qSOFA**) score; or more recently through the use of machine learning models.[121, 127, 167] Most of these efforts focus on the Emergency Department (**ED**) or Intensive Care Unit settings which are data-rich and have a higher prevalence of sepsis compared to the general ward setting.[168-170] However, patients who develop sepsis in the general ward setting have worse outcomes compared to those who develop sepsis in the ED or ICU.[171] Because general ward patients are observed less closely than in the ED or ICU setting with fewer vital signs documented and laboratory tests performed, they represent a proportionally more vulnerable population that could benefit more from an augmented sepsis early warning system. Therefore, the objective of this study was to develop a machine learning model for predicting sepsis in the general ward setting, compare its performance to commonly used instruments for sepsis surveillance such as SIRS and NEWS, and extend the model evaluation using a novel simulated pseudo-prospective trial.[150, 162]

4.4.2. Methods

4.4.2.1. Study Design, Data Sources, and Population

The model was developed and validated using EHR data from Barnes-Jewish Hospital / Washington University School of Medicine in St. Louis, a large, academic, tertiary-care academic medical center. All patients ≥18 years of age that were admitted to the hospital between 1/1/2012 and 6/1/2019 were eligible for inclusion. Patients were excluded if they were admitted to the Psychiatry or Obstetrics services, due to highly variable rates of physiologic data collection. Encounters were excluded if there were no billing code, vital sign, laboratory, service, room, or medication data to indicate a complete patient stay. Encounters were also excluded if the total length of stay was below 12 hours or exceeded 30 days. After assignment of index time and prediction time, further exclusion criteria were applied based on that index time. (**Appendix 35**) To focus on patients most likely to benefit from a risk prediction model, the following populations were excluded: patients who had cultures procured or received antibiotics within 48 hours prior to prediction time; patients who had sepsis present on admission (by admission ICD code); and patients who were in the ICU within 24 hours prior to prediction time. To avoid conflation of post-surgical care with sepsis care, patients were ineligible if they had surgery within 72 hours prior to prediction time. To avoid predicting on patients with excessive missingness, encounters were also required to have at least 3 of each vital sign and at least one complete blood count and one basic or comprehensive metabolic panel test within 24 hours prior to prediction time (**Appendix 36**). This project was approved with a waiver of informed consent by the Washington University in St. Louis Institutional Review Board (IRB #201804121).

4.4.2.2. Sepsis Definition

Sepsis was defined using the Sepsis-3 implementation based on Suspicion of Infection (SOI) determined by concomitant antibiotics and cultures, and the Sequential Organ Failure Assessment (SOFA) score in the ICU setting, and qSOFA (quick SOFA) elsewhere (**Appendix 35**).[127, 172] Anti-infectives for SOI was limited to intravenous anti-infectives except oral vancomycin and metronidazole. In accordance with the Sepsis-3 criteria, SOI required either antibiotics within 72 hours of culture collection, or culture collection within 24 hours of antibiotics.[127] Time of Suspicion of Infection (T_{SOI}) was the earlier of either antibiotic order start time or culture collection time.[172] To meet sepsis criteria, the patient must have had a SOFA or qSOFA score ≥ 2, depending on location, between 48 hours prior to and 24 hours after T_{SOI}. For sepsis cases, time of sepsis onset (T_{Sepsis}) was the same as T_{SOI}.To facilitate model development, each encounter was assigned an index time (T_{Index}), which for sepsis encounters was T_{Sepsis}, and for non-sepsis encounters was 6-hours prior to the maximum of either 1) the midpoint between admission and discharge, or 2) 12 hours into admission. Time of prediction ($T_{Prediction}$) was 6-hours prior to index time (**Appendix 35**).

4.4.2.3. Feature Generation and Engineering

Features were generated from demographic, location, medication, vital sign, and laboratory data available until the time of prediction (**Appendix 32, Appendix 33**). Medications were mapped to classes and subcategories of the Multum MediSource Lexicon by Cerner (Denver, CO). Comorbidities were determined using ICD codes only from prior admissions, and were mapped using the Elixhauser comorbidity system.[107] Time series data were summarized as various univariate statistics (max, mean, etc.) over multiple time horizons (3h, 6h, etc.). Measures of variance such as standard deviation were only computed if there were at least 4 measurements within the time horizon. Missing values, especially results of non-routine lab tests, were likely not missing at random but as a result of clinical judgment, thus were kept as is (**Appendix 34**). Features with > 75% missingness, however, were excluded as they are unlikely to improve performance. For models that required fully non-null input, mean-imputation was used.

4.4.2.4. Model Development

Patient encounters were split at the patient-level to avoid “identity confounding” into train (75%) and test sets (25%).[32] Data transformation parameters were generated based on the training set then applied to both sets. Random search with repeated cross-validation on the training set was used to tune hyperparameters of an eXtreme Gradient Boosting (XGBoost) model, and the optimal combination was used for training on the full training set (**Appendix 38**).[173] Feature importance for the optimized XGBoost model (**XGB opt**) was estimated using the well-validated SHAP approach, a method of credit attribution based on coalitional game theory with useful properties such as additivity and the ability to provide explanations for individual predictions.[96] To condense the model into one that is easier to transport and implement, a “lite” version of the XGBoost model (**XGB lite**) was created using a small subset of features based on the sum of absolute SHAP values across the training set (**Appendix 40**). For comparison, an XGBoost model with default parameters (**XGB unopt**) was trained, as was a logistic regression model with l2 regularization (**LogReg**, **Appendix 41**).

4.4.2.5. Model Performance

The trained models were compared against the Systemic Inflammatory Response Syndrome (**SIRS**) score, National Early Warning Score 2 (**NEWS2**), and quick Sequential Organ Failure Assessment (**qSOFA**) score.[121, 127, 167, 174] Using data from within the 24-hour time window preceding prediction time, SIRS was calculated as the highest score occurring within a 1-hour sliding window; NEWS2 was calculated using the last available measurements; and qSOFA was calculated using the most abnormal measurements. For SIRS, NEWS2, and qSOFA, lack of measurements was interpreted as normal.

Performance of the model was evaluated on bootstrap samples of the test set. Evaluated metrics include (area under) Receiver Operating Characteristic curve (**AUROC**) and (area under) Precision Recall Curve (**AUPRC**). Model calibration was assessed by binning the test set into deciles of predicted risk and comparing their predicted probability of sepsis with actual proportion of sepsis cases. Impact of threshold selection was visualized by plotting performance metrics (specificity, sensitivity, etc.) against the probability threshold.

4.4.2.6. Pseudo-prospective Trial

While the model was trained and evaluated on a single time point per encounter, real world implementation would likely involve continuous risk prediction throughout patient encounters. To better understand the implemented performance of the best performing sepsis prediction algorithm, the model was applied hourly to patient encounters in the test set spanning full admission duration. Patients whose model prediction crossed the threshold maximizing F1 score (harmonic mean of precision and recall) will hereby referred to as having been "alerted on," and for those, "alert time" was defined as the first alert instance for the encounter. For each patient-hour, time-sensitive exclusion criteria (e.g., not in the general ward or already on anti-infectives) were applied again to remove inappropriate alerts. First, the cross tabulation of sepsis status and alert status was generated. Then, among the alerted on, we assessed the proportion of encounters with the following sepsis-related interventions and outcomes: sepsis-relevant culture collection, sepsis-relevant anti-infective administration, ventilator initiation, ICU transfer, sepsis onset, or death.

4.4.2.7. Statistical Analysis

Variables were summarized using frequencies and proportions for categorical data or medians and interquartile ranges (IQR) for continuous data. Statistical comparisons were performed using the Chi-square and Mann-Whitney U tests where appropriate. A p value < 0.01 was considered statistically significant. Analysis and figure generation were performed with Python version 3.7.1 (Python Software Foundation, Beaverton, OR) using the following packages: scipy, numpy, pandas, matplotlib, sklearn, xgboost, and shap.[77, 97-99, 173, 175, 176]

4.4.3. Results

4.4.3.1. Patient Population

From the initial inpatient population of 401,235 encounters, 331,201 met exclusion criteria, leaving 70,034 encounters in the final cohort (**Appendix 36**). Application of the Sepsis-3 criteria identified 2,206 (3.1%) septic patient encounters. Sepsis patients were slightly older (65.6 [56.3 – 74.3] vs. 60.8 [49.4 – 71.2], p < 0.01), more likely to be white (71.3% vs. 61.8%, p < 0.01), had a higher Elixhauser comorbidity score (19 [10 – 29] vs. 9 [1 – 17], p < 0.01), a longer length of stay (12.9 [8.0 – 19.3] vs. 3.9 [2.3 – 6.7], p < 0.01), and higher inpatient mortality (16.6% vs. 0.8%, p < 0.01) (**Table 13**, **Appendix 37**).

Table 13. Cohort characteristics

Variable	Total	Sepsis	Non–sepsis	p[a]
Number of encounters, n (%)	70,034 (100.0%)	2,206 (3.1%)	67,828 (96.9%)	< 0.01*
Age (years), median (IQR)	61.0 (49.6 – 71.3)	65.5 (56.3 – 74.3)	60.8 (49.4 – 71.2)	< 0.01*
Sex (female), n (%)	32,751 (46.8%)	992 (45.0%)	31,759 (46.8%)	0.090
Race, n (%)				< 0.01*
White, n (%)	43,516 (62.1%)	1,573 (71.3%)	41,943 (61.8%)	< 0.01*
Other/unknown, n (%)	3,787 (5.4%)	129 (5.8%)	3,658 (5.4%)	0.378
Black, n (%)	22,285 (31.8%)	487 (22.1%)	21,798 (32.1%)	< 0.01*
Asian, n (%)	446 (0.6%)	17 (0.8%)	429 (0.6%)	0.505
BMI, median (IQR)	27.6 (23.5 – 33.0)	27.2 (23.1 – 33.4)	27.6 (23.5 – 33.0)	0.252
Admitted through ED, n (%)	33,364 (47.6%)	747 (33.9%)	32,617 (48.1%)	< 0.01*
LOS (days), median (IQR)	3.9 (2.4 – 7.0)	12.9 (8.0 – 19.3)	3.9 (2.3 – 6.7)	< 0.01*
Discharge disposition				< 0.01*
Home, n (%)	59,367 (84.8%)	1,185 (53.7%)	58,182 (85.8%)	< 0.01*
Hospice, n (%)	854 (1.2%)	88 (4.0%)	766 (1.1%)	< 0.01*
Acute care facility, n (%)	436 (0.6%)	17 (0.8%)	419 (0.6%)	0.447
Nonacute care facility, n (%)	8,234 (11.8%)	539 (24.4%)	7,695 (11.3%)	< 0.01*
In–hospital death, n (%)	889 (1.3%)	367 (16.6%)	522 (0.8%)	< 0.01*
Other, n (%)	254 (0.4%)	10 (0.5%)	244 (0.4%)	0.589
Sepsis discharge ICD code[b]				< 0.01*
Sepsis, n (%)	1,049 (1.5%)	543 (24.6%)	506 (0.7%)	< 0.01*
Severe sepsis, n (%)	510 (0.7%)	358 (16.2%)	152 (0.2%)	< 0.01*
Septic shock, n (%)	378 (0.5%)	293 (13.3%)	85 (0.1%)	< 0.01*
30–day readmission, n (%)	14,817 (21.2%)	440 (19.9%)	14,377 (21.2%)	0.165
Elixhauser comorbidity score, median (IQR)[c]	9 (1 – 18)	19 (10 – 29)	9 (1 – 17)	< 0.01*

Abbreviations: BMI, body mass index; ED, emergency department; LOS, length of stay; ICD, International Classification of Diseases.

[a] Comparison of variables between sepsis and non-sepsis cohort was performed using Mann-Whitney U test for continuous variables, and χ^2 for categorical variables. Statistical significance ($p < 0.01$) is denoted by *.
[b] Based on sepsis discharge ICD code list from Buchman, Critical Care Medicine, 2020.
[c] Based on Elixhauser comorbidity weights from Moore, Medical Care, 2017

4.4.3.2. Model Performance

The optimized XGBoost model (XGB opt) using all 1,071 features had the highest AUROC (0.862 ± 0.011) and AUPRC (0.294 ± 0.021), compared to the unoptimized XGBoost model (XGB unopt), logistic regression (LogReg), and the lite XGBoost model (XGB lite) all of which had similar performances only slightly worse than XGB opt (**Figure 21**, **Appendix 42**). Scoring systems, however, had significantly lower performance with a loss in AUROC over 0.150.

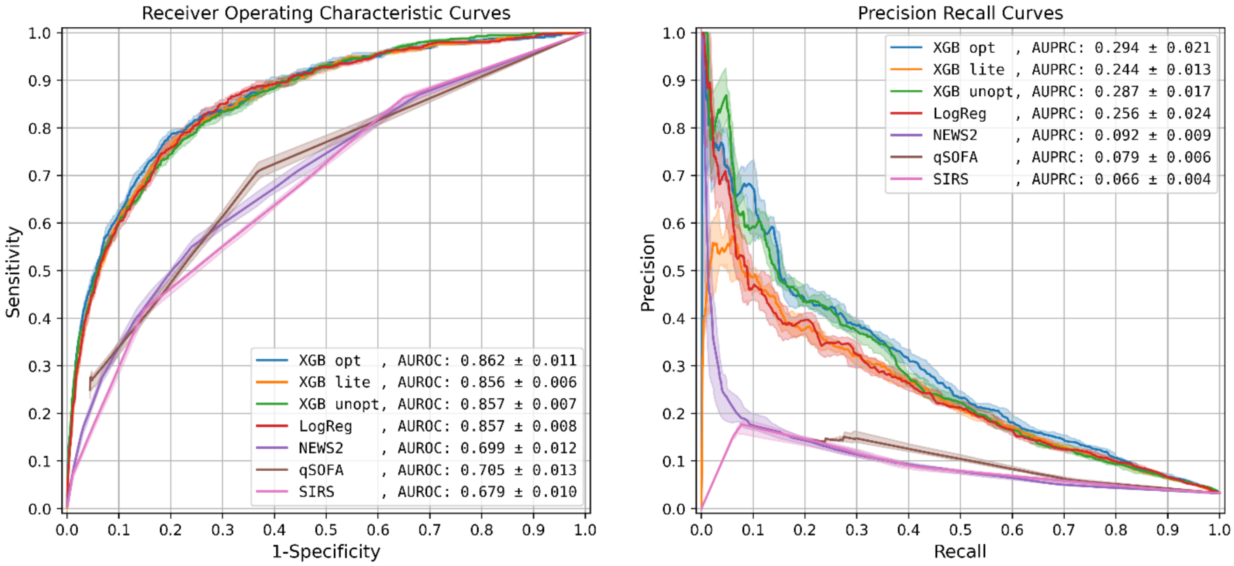
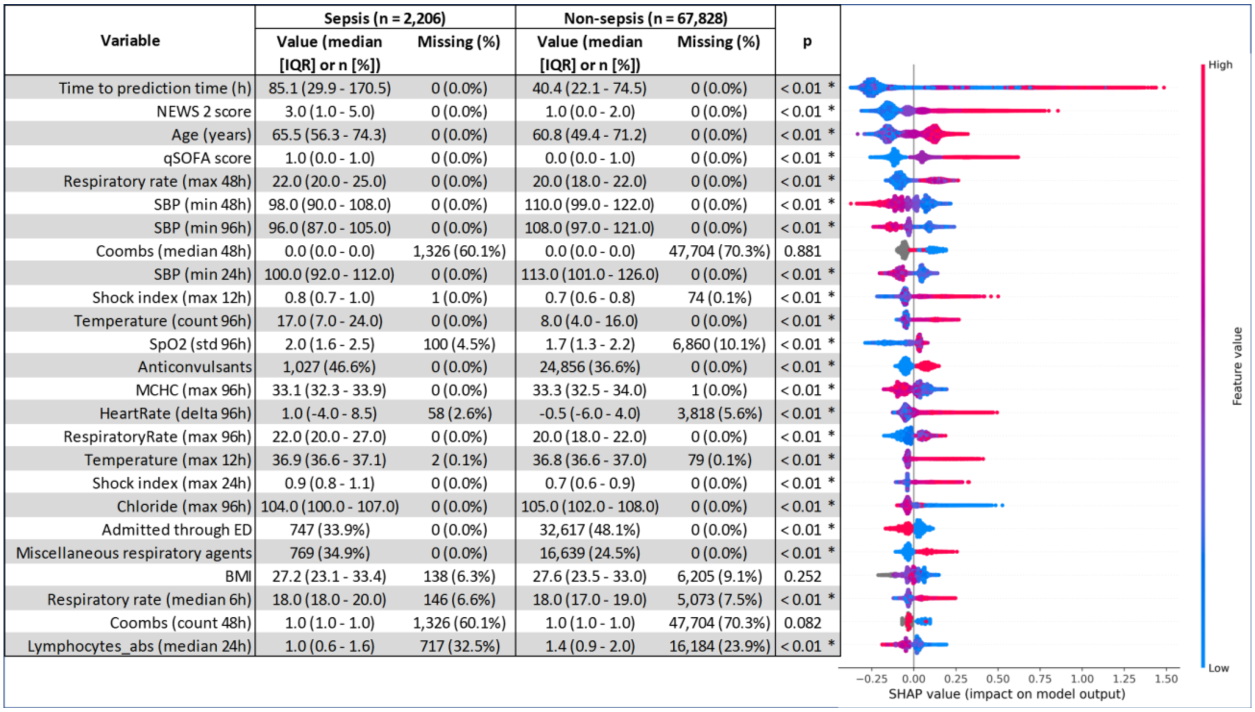


Figure 21. Model Performance: ROC and PR curves

The solid lines represent the 50th percentile curves based on 20 bootstrap (full resampling with replacement) iterations of the test dataset, and the shaded regions represent the area between the 25th and 75th percentiles. Abbreviations: AUROC, area under receiver operating characteristic curve; AUPRC, area under precision recall curve; XGB opt, optimized XGBoost model; XGB lite, simple XGBoost model; XGB unopt, unoptimized, out-of-the-box XGBoost model; LogReg, logistic regression; NEWS2, National Early Warning Score 2; qSOFA, quick Sequential Organ Failure Assessment; SIRS, Systemic Inflammatory Response Syndrome.

The top five most impactful features for the optimized XGBoost model were found to be: time from admission to prediction time, NEWS2 score, age, qSOFA score, and maximum respiratory rate within 48 hours prior to prediction time (**Figure 22**). The calibration curve yielded an r^2 value of 0.837 (**Appendix 42**). The threshold plot demonstrates the tradeoff between precision and recall and revealed highest F1 score (0.346) to be at a threshold of around 0.137 (**Figure 23**).



Variable	Sepsis (n = 2,206)		Non-sepsis (n = 67,828)		p
	Value (median [IQR] or n [%])	Missing (%)	Value (median [IQR] or n [%])	Missing (%)	
Time to prediction time (h)	85.1 (29.9 - 170.5)	0 (0.0%)	40.4 (22.1 - 74.5)	0 (0.0%)	< 0.01 *
NEWS 2 score	3.0 (1.0 - 5.0)	0 (0.0%)	1.0 (0.0 - 2.0)	0 (0.0%)	< 0.01 *
Age (years)	65.5 (56.3 - 74.3)	0 (0.0%)	60.8 (49.4 - 71.2)	0 (0.0%)	< 0.01 *
qSOFA score	1.0 (0.0 - 1.0)	0 (0.0%)	0.0 (0.0 - 1.0)	0 (0.0%)	< 0.01 *
Respiratory rate (max 48h)	22.0 (20.0 - 25.0)	0 (0.0%)	20.0 (18.0 - 22.0)	0 (0.0%)	< 0.01 *
SBP (min 48h)	98.0 (90.0 - 108.0)	0 (0.0%)	110.0 (99.0 - 122.0)	0 (0.0%)	< 0.01 *
SBP (min 96h)	96.0 (87.0 - 105.0)	0 (0.0%)	108.0 (97.0 - 121.0)	0 (0.0%)	< 0.01 *
Coombs (median 48h)	0.0 (0.0 - 0.0)	1,326 (60.1%)	0.0 (0.0 - 0.0)	47,704 (70.3%)	0.881
SBP (min 24h)	100.0 (92.0 - 112.0)	0 (0.0%)	113.0 (101.0 - 126.0)	0 (0.0%)	< 0.01 *
Shock index (max 12h)	0.8 (0.7 - 1.0)	1 (0.0%)	0.7 (0.6 - 0.8)	74 (0.1%)	< 0.01 *
Temperature (count 96h)	17.0 (7.0 - 24.0)	0 (0.0%)	8.0 (4.0 - 16.0)	0 (0.0%)	< 0.01 *
SpO2 (std 96h)	2.0 (1.6 - 2.5)	100 (4.5%)	1.7 (1.3 - 2.2)	6,860 (10.1%)	< 0.01 *
Anticonvulsants	1,027 (46.6%)	0 (0.0%)	24,856 (36.6%)	0 (0.0%)	< 0.01 *
MCHC (max 96h)	33.1 (32.3 - 33.9)	0 (0.0%)	33.3 (32.5 - 34.0)	1 (0.0%)	< 0.01 *
HeartRate (delta 96h)	1.0 (-4.0 - 8.5)	58 (2.6%)	-0.5 (-6.0 - 4.0)	3,818 (5.6%)	< 0.01 *
RespiratoryRate (max 96h)	22.0 (20.0 - 27.0)	0 (0.0%)	20.0 (18.0 - 22.0)	0 (0.0%)	< 0.01 *
Temperature (max 12h)	36.9 (36.6 - 37.1)	2 (0.1%)	36.8 (36.6 - 37.0)	79 (0.1%)	< 0.01 *
Shock index (max 24h)	0.9 (0.8 - 1.1)	0 (0.0%)	0.7 (0.6 - 0.9)	0 (0.0%)	< 0.01 *
Chloride (max 96h)	104.0 (100.0 - 107.0)	0 (0.0%)	105.0 (102.0 - 108.0)	0 (0.0%)	< 0.01 *
Admitted through ED	747 (33.9%)	0 (0.0%)	32,617 (48.1%)	0 (0.0%)	< 0.01 *
Miscellaneous respiratory agents	769 (34.9%)	0 (0.0%)	16,639 (24.5%)	0 (0.0%)	< 0.01 *
BMI	27.2 (23.1 - 33.4)	138 (6.3%)	27.6 (23.5 - 33.0)	6,205 (9.1%)	0.252
Respiratory rate (median 6h)	18.0 (18.0 - 20.0)	146 (6.6%)	18.0 (17.0 - 19.0)	5,073 (7.5%)	< 0.01 *
Coombs (count 48h)	1.0 (1.0 - 1.0)	1,326 (60.1%)	1.0 (1.0 - 1.0)	47,704 (70.3%)	0.082
Lymphocytes_abs (median 24h)	1.0 (0.6 - 1.6)	717 (32.5%)	1.4 (0.9 - 2.0)	16,184 (23.9%)	< 0.01 *

Figure 22. SHAP feature importance

Comparison of variables between sepsis and non-sepsis cohort was performed using Mann-Whitney U test for continuous variables, and χ^2 for categorical variables. Statistical significance ($p < 0.01$) is denoted by *.

Abbreviations: qSOFA, quick sequential organ failure assessment; NEWS2, national early warning system 2; SBP, systolic blood pressure; WBC, white blood cell count; MAP, mean arterial pressure.

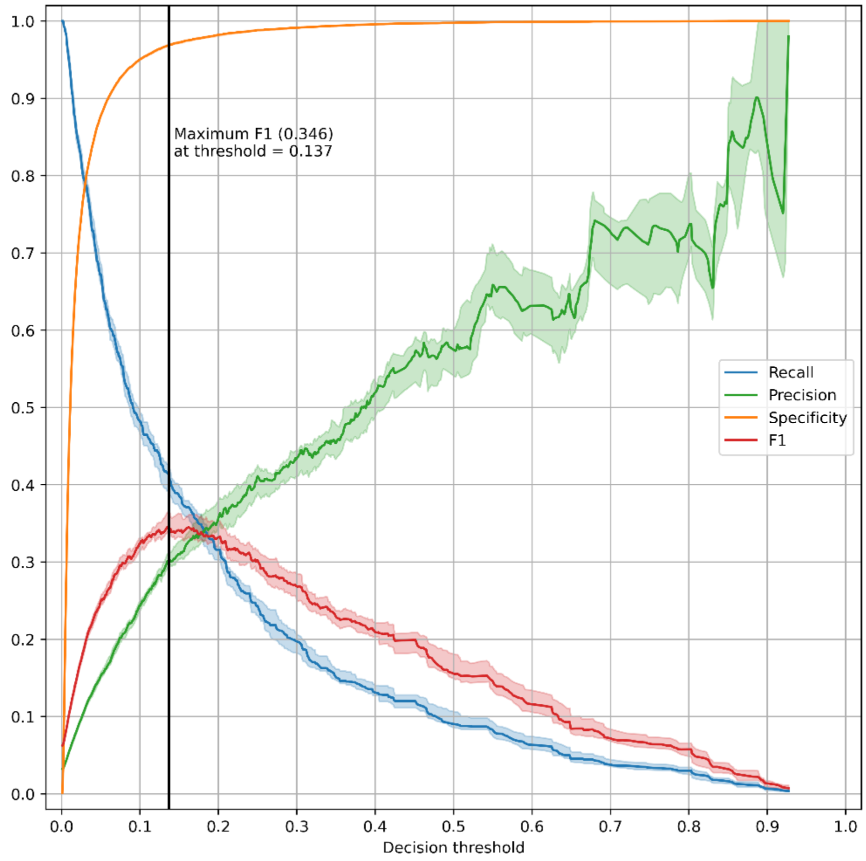


Figure 23. Threshold plot for optimized XGBoost model

The test set was bootstrapped (full resampling with replacement) 20 times and various performance metrics (recall, precision, specificity, and F1) were plotted against threshold value. For each metric, the line and shaded area represents the median and IQR. A vertical black line was drawn at the threshold maximizing F1 score.

4.4.3.3. Pseudo-Prospective Trial

The EHR data of 17,441 encounters in the test set (557 sepsis encounters and 16,884 non-sepsis encounters) were binned hourly into 2,387,482 patient-hours. After exclusions, 3,532 encounters were alerted upon, of which 388 met sepsis criteria (11.0% PPV, **Appendix 42**). Of the 557 sepsis encounters, 388 were alerted upon (69.7% sensitivity). Of the 13,740 non-sepsis encounters, 3,144 were alerted upon (81.4% specificity).

Of the 3,532 alerted encounters, from the time of the first alert, within 48 hours, 23.9% had sepsis-relevant cultures drawn, 13.2% received sepsis-relevant anti-infectives, 2.5% had ventilator initiated, 6.9% experienced sepsis onset, 4.7% were transferred to ICU, and 0.6% died (**Table 14**, **Appendix 45**). Altogether, 29.1% of experienced a sepsis-related intervention or outcome within 48h of first alert.

Table 14. Pseudoprospective trial, outcomes for alerted subjects

Intervention or Outcome	within 24h	within 48h	within 72h
Sepsis-relevant Cultures	600 (17.0%)	843 (23.9%)	1,018 (28.8%)
Sepsis-relevant Anti-infectives	286 (8.1%)	466 (13.2%)	591 (16.7%)
Ventilator Initiation	51 (1.4%)	87 (2.5%)	119 (3.4%)
Sepsis Onset	182 (5.2%)	245 (6.9%)	291 (8.2%)
ICU Transfer	112 (3.2%)	167 (4.7%)	209 (5.9%)
Death	8 (0.2%)	21 (0.6%)	36 (1.0%)
Total	739 (20.9%)	1,028 (29.1%)	1,237 (35.0%)

Of the patients who crossed the set threshold in the pseudoprospective trial, and of those who were not already suspected of or being treated for sepsis, sepsis-related interventions and outcomes within various time horizons were identified.

Visualizations of sample patient trajectories alongside hourly predicted sepsis risk scores facilitated inspections of model successes and failures (**Appendix 46**).

4.4.4. Discussion

The objective of this study was to develop a machine learning model capable of predicting sepsis 6-hours ahead of clinical onset using one of the largest inpatient EHR datasets. Unlike most other sepsis prediction studies which focus on the data-rich ICU or ED setting, this study focused on the general ward setting where the prediction task is made especially challenging due to the sparsity of data and low prevalence.[177] Moreover, the cohort criteria excluded patients who were already suspected of, or were being treated for sepsis, as a clinical prediction model is unlikely to benefit these patients. The resultant cohort represents patients who were not captured by clinical judgment, and thus could benefit from clinical decision support. Further, this study provides a novel way of better estimating real world performance through the assessment of a pseudo-prospective trial.

Excluding patients who were suspected of or were already being treated for sepsis, alongside several other exclusion criteria, resulted in the elimination of the majority of inpatients from the initial population (**Appendix 36**). As a result, the retained sepsis cohort are likely cases of hospital-acquired sepsis or community-acquired sepsis with delayed recognition. Though the restrictive exclusion criteria may limit generalizability, the resultant cohort is more likely to benefit from an automated warning system.

We compared the performance of several machine learning models as well as traditional scoring systems and found the optimized XGBoost to have the best AUROC and AUPRC for detecting sepsis ahead of meeting traditional diagnostic criteria. The "lite" model used on 25 features and had similar diagnostic performance.

Of the important features as determined by SHAP, time from admission to prediction time was the most important, indicating that prolonged length of stay is both a risk factor and outcome for sepsis. qSOFA and NEWS2 scores were also important predictors, demonstrating the utility of these scores as features though deficient on their own. Admission through the ED was associated with a lower probability of sepsis, likely due to the emphasis on sepsis screening in the ED setting. Interestingly, while most medication information was not important for the model, anticonvulsants had a surprisingly high SHAP value with sepsis patients receiving "anticonvulsants" about 10% more frequently than non-sepsis patients (46.6% vs. 36.6%, **Figure 22**). However, the Multum classification for anticonvulsants included medications such as magnesium sulfate and lorazepam which are not always used as anticonvulsants, thus more work is needed on automated feature generation from medication data. Another unexpectedly important feature was the Coombs test, which is unlikely to be related to sepsis, but had noticeably different rates of missingness between sepsis and non-sepsis patients (60.1% for sepsis vs. 70.3% for non-sepsis, **Figure 22**). Comorbidities from prior admissions were noticeably absent from the list of important features, likely because 46.3% of all encounters were first encounters and did not have any prior admissions. It's possible that the importance of comorbidities as features may rise with time, with larger populations with longer histories being collected in the electronic health record and the ability to retrieve information cross-sites.

The pseudo-prospective trial demonstrated a novel approach to better estimating real-world model performance and showed that 29.1% of alerted on patients required sepsis-related intervention or had a sepsis-related outcome within 48h (**Table 14**). While the algorithm was capable of identifying patients who ultimately required cultures (39.0%) and anti-infectives (28.1%), the actual incidence of Sepsis-3 onset after patients were alerted on was relatively low (11.0% at any point after and 6.9% within 48h). This may be due to problems in labeling – despite our attempt to exclude surgical patients from the cohort, they are not capable of being excluded in a prospective basis, and frequently met sepsis criteria. Moreover, alerted on patients may be critically ill and treated for sepsis but not meet Sepsis-3 criteria. Also, many patients who present to the ED have higher scores which improve through interventions, but then have scores that rise again later during the hospital course. Since only the first time a patient crossed the sepsis threshold was evaluated here, the subsequent and potentially more important clinical changes would be missed. The pseudo-prospective trial highlights some of the anticipated challenges of translating a diagnostic scoring method from a retrospective data set to a prospective population, which necessitates further investigation.

Impressively, the unoptimized XGBoost solution had a median AUROC just 5% lower than the optimized version, and similar performance to the optimized logistic regression model and the lite XGBoost model. The relatively small benefit conferred by the more complex model compared to logistic regression is consistent with prior literature.[160] If the added complexity is problematic – for interpretability, debugging, or implementation – then it could be argued that the simpler logistic regression model is preferred despite the performance loss. Though NEWS2 and qSOFA were very important features in XGBoost, the gap between traditional scoring systems and machine learning models was noticeable with the worst ML model conferring a 15.1% AUROC improvement over the best traditional scoring system.

This study has limited generalizability as a single institution study. The study used an interpretation of the Sepsis-3 definition and is likely to generalize poorly to sites using alternate definitions.[118, 127, 172] By design, the study was focused on the general ward setting, and the results are not applicable to other settings. Many of the excluded subpopulations (children, surgical, etc.) warrant further investigation. While a pseudo-prospective trial was performed, a true prospective study is needed to gauge real-world performance. The pseudo-prospective trial could be further improved by investigating repeated alerts, incorporating alert lock-out periods, accounting for measurement-to-documentation time gap, etc. For the pseudo-prospective trial, a threshold was assigned to maximize the F1 score. However, further work is necessary to define an operationally meaningful threshold. For the calculation of the qSOFA score, GCS was missing in our dataset and assumed normal, which may negatively impact the sepsis label assignment process. However, Seymour et al. found that the lack of GCS in the VA dataset did not significantly reduce the predictive validity of qSOFA.[127] As is typical of studies using electronic health records data, there were and likely remain problems concerning missingness and accuracy of clinical data.

4.4.5. Conclusion

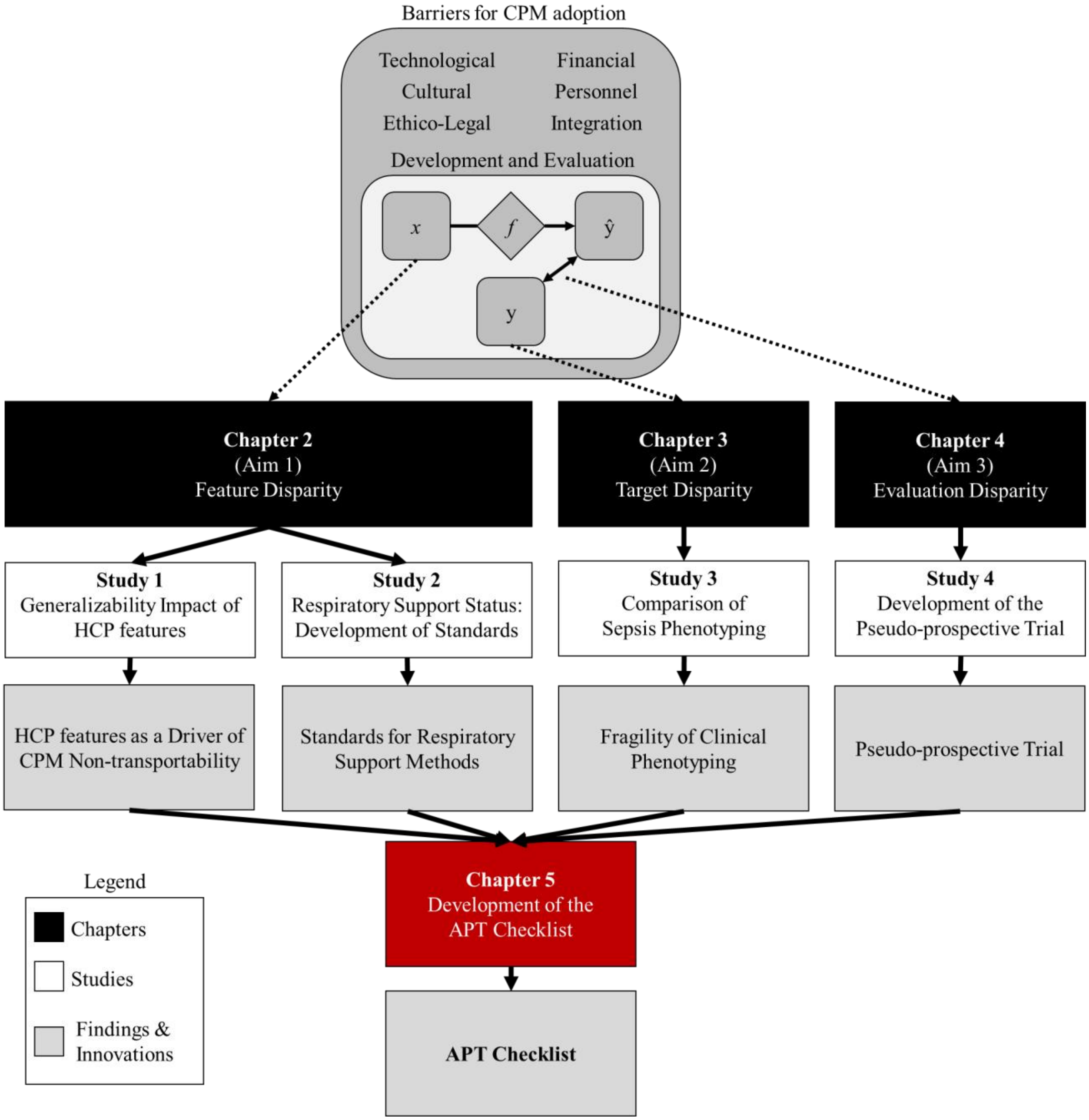
A machine learning model designed to predict sepsis 6-hours ahead of meeting diagnostic criteria yielded an AUROC of 0.862 ± 0.011 and AUPRC of 0.294 ± 0.021. Pseudo-prospective evaluation of the model meaningfully expanded the understanding of model performance, and revealed relatively good clinical performance, despite a large class imbalance.

4.5. Discussion and Conclusion

The aim of this chapter was to identify, assess the impact of, and propose solutions for the disparity between the evaluation design conducted by the CPM development organization and the expected implementation behavior of the CPM-based CDS in the adopting organization, resulting in incomparable performance metrics and confusing the assessment of CPM transportability, thus limiting CPM transportability. Implementation behavior includes temporality of CPM executions as well as those specific to CPM-based CDS such as alert snoozing behavior. These factors are rarely considered *in silico* yet can have critical impact on the success of CPM-based CDS *in vivo*, thus must be considered *in silico* as well. To provide a framework for uniting and extending the individual and independent efforts in incorporating these implementation factors into CPM evaluation design, we proposed the novel pseudo-prospective trial, a DES-like simulation framework for assessing CPM behavior over time which allows for the incorporation of various, time-varying implementation factors into the *in silico* evaluation process. It's found that the pseudo-prospective trial significantly enhances and extends CPM performance understanding, thus shows promise as a tool for bringing parity between the CPM evaluation design and the expected implementation behavior of the CPM-based CDS. Organizations seeking to adopt an externally developed CPM and implement as a CDS should first identify how a CPM is likely to implemented as a CDS in their workflow, and assess parity between expected implementation behavior and the evaluation design carried out by the CPM development team. If significant disparities are present, the adopting organization must understand that the performance metrics reported by the CPM development team cannot be interpreted directly, and if possible, request a re-evaluation that does take into account the implementation factors based on expected CPM-based CDS behavior.

Chapter 5. Development of the APT Checklist

5.1. Introduction



Barriers for CPM adoption
Technological
Financial
Cultural
Personnel
Ethico-Legal
Integration
Development and Evaluation
x
f
ŷ
y
Chapter 2
(Aim 1)
Feature Disparity
Chapter 3
(Aim 2)
Target Disparity
Chapter 4
(Aim 3)
Evaluation Disparity
Study 1
Generalizability Impact of HCP features
Study 2
Respiratory Support Status: Development of Standards
Study 3
Comparison of Sepsis Phenotyping
Study 4
Development of the Pseudo-prospective Trial
HCP features as a Driver of CPM Non-transportability
Standards for Respiratory Support Methods
Fragility of Clinical Phenotyping
Pseudo-prospective Trial
Chapter 5
Development of the APT Checklist
APT Checklist
Legend
Chapters
Studies
Findings & Innovations

Figure 24. Chapter 5 Overview

The objective of this chapter is to describe the development of the **A**ssessment of Clinical **P**rediction Model **T**ransportability (**APT**) Checklist. As described in chapter 1, external implementation of CPMs or conversely, adoption of externally developed CPMs has numerous potential benefits such as the de-duplication of CPM development efforts, enabling of low-resource organizations to participate in the usage of modern CPMs, facilitating external validation studies, and overall, encouraging collaborative development of CPMs. While these efforts have been hindered by the lack of health IT interoperability in the past, modern efforts in policy, standards, and tools increasingly ease the adoption of externally developed CPMs. However, naïve implementation of externally developed CPMs can result in significantly degraded performance or lack of transportability. Thus, the assessment of CPM transportability is critically important in enabling this new paradigm of open sharing and adoption of CPMs. While there is a wealth of research on machine learning model generalizability and transportability including subjects such as data drift, methods for estimating external validity, and model updating methods; and while there are numerous frameworks guiding the development, evaluation, reporting, and systematic review of CPMs; there are no frameworks for assessing the transportability of CPMs. So, the objective of this dissertation and this chapter is simply to address this gap by proposing a novel framework for the assessment of CPM transportability. Broadly, the success of CDS based on an externally developed CPM can be impacted by a wide variety of factors such as resource availability – human, technical, financial, or otherwise – or the culture surrounding use of ML/AI at the adopting organization. For the purpose of this dissertation, the scope was limited to matters concerning the development of CPMs. To that end, first, extant CPM-associated frameworks were synthesized to identify the key dimensions of CPM development, finding three domains to be fruitful targets of interrogation: feature, target, and evaluation; each comprising a specific aim and chapter of the dissertation. Altogether, four studies were conducted as part of these chapters and aims, each identifying, characterizing, and/or propose solutions for challenges to transportability of CPMs. The findings and innovations of these studies – as described in chapters 2, 3, and 4 – are integrated into and merged with the synthesis of extant CPM-associated frameworks, resulting in the novel APT checklist. While rigorous evaluation of the framework remains future work, the checklist shows promise as a tool for CPM adopting organizations to evaluate the transportability of candidate CPMs.

5.2. Overview

Structurally, this chapter will begin by reiterating the motivating objectives as has already been done in the preceding section, followed by a background section on the following topics – frameworks in biomedical informatics and barriers to success of CPM-based CDS. Then, the methods and results of this chapter are presented including the synthesis of extant CPM frameworks, findings and innovations from each study conducted as part of this dissertation, and the integration of the latter into the former, culminating in the development of the APT checklist.

5.3. Background and Significance

5.3.1. Frameworks in Biomedical Informatics

There are many different ways to approach a critical problem in the domain of biomedical informatics such as CPM transportability. One such way is to, for example, focus on a narrow problem within CPM transportability, developing and comparing methods for CPM updating in response to data drift.[38] Another broader approach for addressing biomedical informatics problems, which has been prescribed by the board of the American Medical Informatics Association as a core competency as a fundamental scientific skill for students of the discipline, is the creation of novel frameworks.[178] This dissertation addresses the biomedical informatics problem of CPM non-generalizability through the development of a novel framework in the form of a checklist.

While a single universal definition is elusive, especially in comparison to the idea of theories or models, frameworks are tools created to “characterize, describe, guide, analyze, and evaluate phenomena and processes.” [179] Frameworks can be used to predict and explain phenomena or describe and guide practice, can take the form of checklists, and can be evaluative.[180] Given the diversity of problems that are the subject of biomedical informatics research, there is an accompanying diversity of theories, frameworks, and tools employed or developed by researchers and practitioners of biomedical informatics. Many relate to the adoption or implementation of information technologies in healthcare – the APT checklist proposed in this dissertation is certainly associated to frameworks on technology acceptance or implementation science.[180-183] While there are no frameworks specifically for the assessment of CPM transportability, there are those in the vicinity for the purposes of CPM development, comparative evaluation, reporting, and systematic reviews such as TRIPOD, CHARMS, PROBAST, and GRASP.[43-46] The APT checklist draws and builds on these frameworks as described in the methods and results section of this chapter. Through the interrogation of literature – including those describing frameworks – on the implementation of CPM-based CDS, concerns beyond those addressed by this dissertation were surfaced, which is the subject of the following section.

5.3.2. Barriers to Success of CPM-based CDS

As described in chapter 1, the focus of this dissertation is on CPM transportability – whether a CPM could maintain commensurate performance and clinical utility on completely external populations. However, CPMs are rarely products in and of themselves to be used directly by end-users, but rather require packaging into a CDS and incorporated into clinical workflows.[184] As such, success of CPM depends on a constellation of factors beyond those specific to the development and evaluation of CPMs or the subject of this dissertation (**Figure 4**). So, this section is intended to provide the background on these other critically important factors for the broader consideration of transportability, the clinical utility of CDS based on externally developed CPMs.

Technological – Akin to McDermott's concept of technical reproducibility, technological factors relate to the capacity of the adopting organization to faithfully re-implement the model.[24] While simple linear regression models can be easily transported by publishing or otherwise providing the coefficients and intercept, more complex models may require additional work including directly sharing a frozen model, controlling the version of all dependencies, etc. Interoperability efforts, code sharing, proper documentation, and containerization can all help mitigate or overcome technological barriers to CPM adoption.

Financial – Implementation, CDS follow-up, and maintenance all take non-trivial resources, especially if the infrastructure needs to be constructed rather than repurposed. Thus, CPM adoption efforts require institutional support including financial support. Ideally, the CPM-based CDS would be self-sustaining by encouraging behavior that is money-saving for the health system such as reducing readmissions. However, the direct return on investment may be difficult to ascertain, only be realized long-term, or worse, may actually be negative even while improving patient care (e.g., reducing unnecessary but lucrative tests). To reduce the likelihood of project stalling or abandonment, the CPM adoption effort should be aligned with the vision of the institution and their funding mechanisms, which may be doubly aligned with national quality improvement programs with financial incentives.[21]

Cultural – The landscape of opinion on CDS, especially those involving predictive models using ML/AI, is complex. On one hand, the excitement of augmenting clinical decision support using modern computational hardware and software still burns bright. However, the imposition of EHR adoption, coupled with usability issues, increased documentation load, and overuse of uninformative interruptive alerts have resulted in frustration with implemented CDS systems.[185] As such, additional efforts likely to result in more disruptive alerts are understandably met with weariness and skepticism. Further, there is resistance or even disdain for CDS perceived as overbearing due to the underlying presumption of a "cookbook medicine" framework, which is seen as taking autonomy and joy away from the practice of medicine.[186] As such, the CPM adopting organization may not be culturally prepared for additional CPM-based CDS tools. Further, there may be subcultures within the institution – divided by specialty, role, age, etc. – which may be more or less receptive to CPM-based CDS. Thus it is critical to understand the culture of the institution and its stakeholders prior to CPM adoption.

Personnel – Implementing, debugging, updating, reporting, and potentially performing further testing such as A/B testing or running prospective impact trials require varied and skilled workforce ranging from computer scientists, biomedical informaticists, to clinical subject matter experts as well as project managers, human-computer-interface design experts, statisticians, EHR-proficient software engineers, and so on. Further, the CPM-based CDS workflow may involve follow-up by specialized personnel such as rapid response nursing which may require staffing up. In sum, insufficient human resources could hinder CPM adoption.

Ethico-Legal – First, the regulatory landscape of CPM as laid out by the FDA is still-evolving. While an in-depth discussion is beyond the scope of this discussion, the FDA is developing a framework for handling AI/ML-based Software as Medical Device (**SaMD**). Of note, one critical point of contention is the handling of adaptive models – CPM adoption may involve model fine-tuning and/or updating which would change the model such that it is no longer the same as the explicitly approved model.[187] Thus, ensuring that the CPM adoption effort conforms to regulatory standards as they evolve can be potentially challenging. Second is the concern of culpability when injury occurs based on faulty recommendations by CPMs. The current legal perspective is based on standard of care, but long-term incorporation of CPMs in clinical practice may make CPM usage as part of standard of care.[188] While legal liability falls almost entirely on the provider, researchers have observed what has been termed automation bias in which providers overly trust recommendations from automated CDS. Thus, the stakeholders of the CPM adoption effort must be prepared to agree on, then shoulder the burden of ethical if not legal responsibilities of an active CPM-based CDS. Third and last is the issue of unfairness. Often CPMs are trained in high-resource, academic medical centers whose case-mix may be significantly different from the overall demographic, thus resulting in poor model performance for low-resource health systems or minority populations, which could be seen as exacerbating the inequality problem in the US healthcare system. Identifying and reducing bias in CPMs is an active field of research (the details of which will be omitted here) which should be considered during a CPM adoption effort.[189]

Integration – Most CPM efforts are conceptualized as eventually translating into CDS in the form of interruptive alerts as they are most active form of decision support, demanding clinician attention, resulting in action or inaction that can be easily logged and analyzed.[190] Many part of this manuscript presupposes CPM adoption efforts as alert implementation efforts. However, CDS can be more extreme by completely prohibiting clinicians from taking certain actions, or be more lax by simply proving the user with information with no incentives in either direction or requirement for follow-up.[184] Further, there are many additional modifiable aspects of CPM-based CDS in the integration into clinical workflow including the CDS user experience/interface design, intended user base, involved stakeholders, incentivization, follow-up action choices, and so on. Even when just talking about interruptive alerts, as discussed in chapter 4, they can have modifiable "snooze" periods, triggering action, triggering provider criteria, threshold selection, and so on. All of these integration factors can significantly impact CPM adoption success, thus must be carefully chosen.

5.4. Methods

5.4.1. Synthesis of Extant CPM-Associated Frameworks

The first step in the development of the APT checklist was a review and synthesis of extant CPM-associated frameworks to determine the dimension of CPM development and evaluation that warrant investigation as to their impact on CPM transportability. To this end, four CPM frameworks were reviewed – TRIPOD, CHARMS, PROBAST, and GRASP.[43-46] To enable the synthesis of these frameworks, each framework was converted to, if not already in the form of, a labeled list of items (**Appendix 1, Appendix 2, Appendix 3, Appendix 4**). This was then used to synthesize the frameworks into a single framework outlining the major categories of concerns pertaining to CPM development and evaluation: background, population, target, modeling, evaluation, and validation (**Table 1**). Three of those (sub)-categories were determined to be fruitful targets of further investigation: target, feature, and evaluation. The base framework constructed through the synthesis of prior frameworks was then enriched through the findings and innovations discovered or proposed through the four studies that were conducted as part of this dissertation.

5.4.2. Findings and Innovations from Aim 1 (Chapter 2)

The objective of chapter 2, or specific aim 1, was to identify, characterize, and propose solutions for challenges to transportability of CPMs, specifically focusing on features used as input by CPMs. The first study found a class of features that are heavily influenced by site-specific factors such as hospital processes, documentation culture, and choice of hardware/software, termed HCP features as opposed to features based on the patient's pathophysiology. These HCP features were found to contribute to the overfitting of CPMs, deceptively improving internal and estimated external performance at the cost of hurting actual external performance. Thus, the implication for the APT checklist and the recommendation for organizations seeking to adopt a CPM is to assess the presence and reliance of HCP features, and if heavily reliant, abandon or plan to update or retrain. In the second study, the insufficient coverage of EHR data by standards is proposed as a cause of CPM feature heterogeneity, and respiratory support methods was identified as an area of EHR data lacking standards. A novel classification system and an accompanying set of EHR-agnostic heuristics was proposed to mine respiratory support episode information from EHR data in a standardized manner. The implication for the APT checklist and the recommendation for organizations seeking to adopt a CPM is to check if features are based on standards, and if not, check for the documentation of ad-hoc standards or at the least, a coherent method of mapping from raw data to features. The recommendations based on the findings and innovations of the studies are integrated into the APT checklist as described in a later section.

5.4.3. Findings and Innovations from Aim 2 (Chapter 3)

The objective of chapter 3, or specific aim 2, was to identify, characterize, and propose solutions for challenges to transportability of CPMs, specifically focusing on features used as the prediction target by CPMs. For the study conducted as part of chapter 3, sources of target heterogeneity were stratified into two levels – the macro level having to do with disease definition and understanding, and the micro level relating to the specifics of the phenotyping criteria. For a complex, syndromic condition without a gold-standard diagnostic test commonly used as a target for CPMs – sepsis – we found significant heterogeneity in both the macro level as well as the micro level, resulting in significant heterogeneity in the sepsis cohorts' patient characteristics as well as clinical outcomes. The implication for the APT checklist and the recommendation for organizations seeking to adopt a CPM is to assess agreement with the phenotyping approach used by the developers, both on a macro, definition-level and also on the micro, criteria-level. The recommendations based on the findings and innovations of the studies are integrated into the APT checklist as described in a later section.

5.4.4. Findings and Innovations from Aim 3 (Chapter 4)

The objective of chapter 4, or specific aim 3, was to identify, characterize, and propose solutions for challenges to transportability of CPMs, specifically focusing on evaluation design of CPMs. In the study conducted as part of chapter 4, disparities between the design of CPM evaluation performed by the developers and the expected behavior of CPM-based CDS once implemented, are hypothesized to impede the assessment of CPM transportability. For example, alert silencing behavior is common in CPM-based CDS, but rarely integrated into the evaluation process. So, a novel simulation-based evaluative framework was proposed, termed the pseudo-prospective trial, which can facilitate the integration of these additional factors into the CDS evaluation process. It was found that the pseudo-prospective trial shows promise as a foundation for bringing parity between the CPM evaluation design and the expected implemented behavior of the CPM-based CDS. The implication for the APT checklist and the recommendation for organizations seeking to adopt a CPM is to assess the (dis)parity between the design of the evaluation carried out by the CPM development team and the expected behavior of the CDS based on the CPM once implemented such as through the pseudo-prospective trial. The recommendations based on the findings and innovations of the studies are integrated into the APT checklist as described in a later section.

5.4.5. Assembly into the APT Checklist

As previously described, the foundational framework was developed through the synthesis of the extant CPM-associated frameworks. The framework was then enriched with the recommendations based on the findings and innovations of the three chapters of the dissertation, each corresponding to one specific aim, all together comprising of four studies. Categories deemed out of scope for the dissertation were enriched through literature review as discussed in chapter 1.

5.5. Results

The APT checklist is a biomedical informatics framework intended to guide the assessment of CPM transportability, aimed toward those seeking to adopt an externally developed CPM (**Table 15**). While potentially applicable to other models, this proposed checklist is primarily intended for the acute care setting, and for supervised, binary classification, machine learning models. The APT checklist is comprised of 6 categories: background, population, target, modeling, evaluation, and validation; and contains 17 items in total. Of those items relevant to CPM development and evaluation, 15 were found to have implications for CPM transportability, and thus recommendations for each have been documented as a separate column for the checklist. Some of those recommendations are directly based on those identified through the conducting of the four studies contained in this dissertation. The remaining recommendations are based on prior literature, which some based on specific articles. The APT Checklist is as follows:

Table 15. The APT Checklist

Category	Description	Transportability
Background	Study rationale, scope, purpose, use-case	Alignment in vision for CPM usage
Population	Data source and study design (RCT, registry, etc.) Study setting and period	Similar study setting, case mix, and proximity in time period
	Inclusion/eligibility criteria	Ability to execute eligibility criteria
	Population characteristics	Model transportability metrics (univariate or joint[a])
Target	Outcome/target definition	**Parity of outcome concept, definition, and criteria, as well as ramifications of disparity**[b]
Modeling	Predictor descriptions (type, what, when, etc.)	Availability and syntactic parity of features
	Missing data analysis and handling	**Usage and parity of Health Care Process (HCP) features**[c]
	Predictor manipulation/feature engineering	**Feature mapping to standards where available, otherwise specification of ad-hoc standards**[d]
	Model type	Technological reproducibility
	Model training procedure including feature selection	
	Model updating/recalibration	Capacity and procedure to update/recalibrate[e]
Evaluation	Model evaluation procedure	Parity between evaluation procedure and intended use case as by, e.g. **Pseudoprospective trial results**[f]
	Interpretation of results	
Validation	Extent of validation (internal, external)	Investigate prior assessments of external validity
	Usability	
	Impact (clinical effectiveness, patient safety, healthcare efficiency)	

[a] Using e.g. adjMMD from Song (Nat Comm, 2020)
[b] Investigated in chapter 3, study 3 of this dissertation
[c] Investigated in chapter 2, study 1 of this dissertation
[d] Investigated in chapter 2, study 2 of this dissertation
[e] Using, e.g., approach by Davis (JAMIA, 2019)
[f] Investigated in chapter 4, study 4 of this dissertation

Chapter 6. Summary and Conclusions

6.1. Summary

Modern interoperability efforts in policy, standards, and tools increasingly facilitate external implementation of CPMs. Naïve external implementation, however, is prone to failure, resulting in significantly degraded performance when implemented. While there are ongoing research in ML model generalizability and while there numerous CPM-associated frameworks proposed to guide the development, comparative evaluation, and systematic reviews of CPMs, there are no frameworks designed to guide the assessment of CPM transportability. Thus, the objective of this dissertation was to address this critical gap in literature through the proposal of a novel framework specifically for the evaluation of CPM transportability. To that end, prior CPM-associated frameworks were synthesized and reviewed, finding three categories worthy of targeted investigation – disparities in feature, target, and evaluation – each comprising a specific aim and chapter of the dissertation. In the three chapters, four studies were conducted, each identifying, assessing the impact of, and/or providing solutions for the barriers to CPM non-transportability. Finally, recommendations based on the findings and innovations of the studies were assembled into the APT checklist, the primary innovation and contribution of this dissertation.

6.1.1. Aim 1

The objective of aim 1, chapter 2, was to identify, assess the impact of, and/or provide solutions to barriers of CPM transportability, focusing on the features used as inputs for CPMs. To that end, two studies were performed, the first of which identified the impact of HCP features on external generalizability, finding that HCP features in fact reduce external generalizability compared to its foil, PP features. Implications for the APT checklist is that CPMs heavily reliant on HCPs must be avoided or there must be a plan to retrain. The second study, based on the idea that the lack of standards contribute to feature heterogeneity, proposed a novel classification system and heuristics for a section of EHR data previously not mapped by standards. Implications for the APT checklist is that for any given CPM, the path from raw data to features must be unambiguous and readily reproducible, ideally based on widely accepted standards, and if not, based on ad hoc standards or through transparent documentation.

6.1.2. Aim 2

The objective of aim 1, chapter 2, was to identify the causes and characterize the impact of heterogeneity in labels required for CPM development and its ramifications on CPM transportability. A study was conducted evaluating the variability in sepsis definitions, phenotyping criteria, and cohort characteristics – finding significant heterogeneity in all levels, highlighting the fragility of such clinical phenotyping approaches. Implications for the APT checklist is that differences between target labeling approaches must be identified and assessed if significant enough to warrant abandonment, re-evaluation, or re-training.

6.1.3. Aim 3

The objective of aim 1, chapter 2, was to characterize and provide solutions for heterogeneity in the framing of CPM evaluation approaches by bridging the gap between CPM evaluation design and expected implemented behavior of CPM-based CDS. Identifying the gap between CPM evaluation and expected implementation behavior of CPM-based CDS, a study was conducted proposing and demonstrating a novel evaluation framework termed the pseudo-prospective trial.

Implications for the APT checklist is that disagreements on how a CPM was evaluated and how it's intended to be used can make irrelevant the evaluation results including measures of performance. First, those disagreements must be identified, and if deemed significant, the CPM in question must be re-evaluated with the implementation behavior factors integrated into the design.

6.2. Findings and Innovations

The primary and overall contribution of this dissertation is the proposal of a novel framework called the APT Checklist for guiding CPM adoption through the evaluation of CPM transportability. Secondary findings and innovations include those from individual studies within the chapters based on specific aims, chapters 2, 3, and 4. In study 1, HCP variables were found as a driver of CPM non-transportability. In study 2, the fragility of clinical phenotyping often used to identify prediction targets for CPMs was found in the context of sepsis. In study 3, a novel classification system and meta-heuristics for a section of EHR data previously lacking in standards – respiratory support methods – was proposed and evaluated. Finally, in chapter 4, a novel CPM evaluation design termed the pseudo-prospective trial was developed and demonstrated using sepsis prediction in the general ward setting as the clinical backdrop. Overall, this work contributes to the body of biomedical informatics literature guiding the success of informatics interventions.

6.3. Contribution to Informatics

Comparative evaluation of biomedical informatics solutions is critical to the success and advancement of the discipline. However, despite the proliferation of CPMs and the increasing ease of cross-site implementation, there are no frameworks guiding the evaluation of CPM transportability. This dissertation addresses this critical lack in literature by proposing a novel framework for CPM transportability intended for use by those seeking to adopt externally developed CPMs. While further iteration and evaluation remains future work, the checklist shows promise as a tool for reducing the risk of adopting non-transportable CPMs.

6.4. Future Directions

First and foremost, rigorous meta-evaluation of the APT checklist remains future work. Overall, the evaluation should focus on assessing whether adhering to the recommendations made by the framework indeed reduces CPM non-generalizability across multiple settings, compared to other competing frameworks. In addition, further iteration and refine of the APT checklist can be performed through merging with frameworks in the non-clinical ML discipline, by regular reviews of the CPM literature, or by conducting studies targeting high-yield problems in CPM transportability. Also, since the APT checklist was developed primarily for binary classification ML models using EHR data to be used in the acute care setting, the scope of inquiry could be expanded to include other types of CPMs used in healthcare. Finally, since the APT checklist was intended for use by those seeking to adopt an externally developed CPM, the checklist could be re-interpreted as a didactic tool for CPM developers instead.

6.5. Conclusion

The overall objective of this dissertation was to address a critical gap in literature through the proposal of a novel framework for the evaluation of CPM transportability. To that end, four extant CPM-associated frameworks – TRIPOD, CHARMS, PROBAST, and GRASP – were synthesized and reviewed, finding the following domains worthy of further, targeted investigations: feature, target, and evaluation. For these three specific aims, four studies were conducted in total, each addressing disparities between the CPM development site and the external implementation site that give rise to CPM non-transportability. Recommendations based on the findings and innovations from these studies were then incorporated into the synthesis of prior CPM frameworks, resulting in the novel APT checklist, shows promise as a as a tool for reducing the risk of adopting non-transportable CPMs.

References of Included Studies

Study 1 is in preparation for publication and has not yet been published. Study 2 has been published in the Journal of American Medical Informatics Association in 2022.[191] Study 3 has been published in Critical Care Medicine in 2021.[192] Study 4 has been published in the Frontiers of Digital Health in 2022.[193]

References

1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? Bmj. 2009;338.

2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical care medicine. 1985;13(10):818-29.

3. Blumenthal D. Wiring the health system—origins and provisions of a new federal program. New England Journal of Medicine. 2011;365(24):2323-9.

4. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25:1097-105.

5. Adibi A, Sadatsafavi M, Ioannidis JP. Validation and utility testing of clinical prediction models: time to change the approach. Jama. 2020;324(3):235-6.

6. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MM, Dahly DL, Damen JA, Debray TP. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj. 2020;369.

7. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. bmj. 2019;367.

8. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C. Prediction models for cardiovascular disease risk in the general population: systematic review. bmj. 2016;353.

9. Mateen BA, Liley J, Denniston AK, Holmes CC, Vollmer SJ. Improving the quality of machine learning in health applications and clinical research. Nature Machine Intelligence. 2020;2(10):554-6.

10. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innovations. 2020;6(2).

11. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ digital medicine. 2019;2(1):1-5.

12. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, de Ridder MA, Seinen TM, Williams RD, Rijnbeek PR. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. Journal of the American Medical Informatics Association. 2022;29(5):983-9.

13. Mandl KD, Kohane IS. No small change for the health information economy. The New England journal of medicine. 2009;360(13):1278.

14. Dolin R, Boxwala A, Shalaby J. A pharmacogenomics clinical decision support service based on FHIR and CDS hooks. Methods of information in medicine. 2018;57(S 02):e115-e23.

15. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu L-M, Moons KG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC medical research methodology. 2014;14(1):1-11.

16. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of clinical epidemiology. 2015;68(1):25-34.

17. Johnson AE, Pollard TJ, Naumann T. Generalizability of predictive models for intensive care unit patients. arXiv preprint arXiv:181202275. 2018.

18. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, Pestrue J, Phillips M, Konye J, Penoza C. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine. 2021;181(8):1065-70.

19. Matics TJ, Sanchez-Pinto LN. Adaptation and validation of a pediatric sequential organ failure assessment score and evaluation of the sepsis-3 definitions in critically ill children. JAMA pediatrics. 2017;171(10):e172352-e.

20. Irschik S, Veljkovic J, Golej J, Schlager G, Brandt JB, Krall C, Hermon M. Pediatric Simplified Acute Physiology Score II: Establishment of a New, Repeatable Pediatric Mortality Risk Assessment Score. Frontiers in pediatrics. 2021;9.

21. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, Nderitu N, Poon EG. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA open. 2020;3(2):167-72.

22. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Affairs. 2014;33(7):1148-54.

23. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006;144(3):201-9.

24. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Science Translational Medicine. 2021;13(586).

25. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science. 2000;46(2):186-204.

26. Eccles MP, Foy R, Sales A, Wensing M, Mittman B. Implementation Science six years on—our evolving scope and common reasons for rejection without review. Springer; 2012.

27. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine. 1999;130(6):515-24.

28. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal. 2014;35(29):1925-31.

29. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset shift in machine learning: Mit Press; 2008.

30. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S. The clinician and dataset shift in artificial intelligence. The New England journal of medicine. 2021;385(3):283.

31. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.

32. Neto EC, Pratap A, Perumal TM, Tummalacherla M, Snyder P, Bot BM, Trister AD, Friend SH, Mangravite L, Omberg L. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. NPJ Digital Medicine. 2019;2(1):1-6.

33. Cerqueira V, Torgo L, Mozetič I. Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning. 2020;109(11):1997-2028.

34. de Jong VM, Moons KG, Eijkemans MJ, Riley RD, Debray TP. Developing more generalizable prediction models from pooled studies and large clustered data sets. Statistics in medicine. 2021;40(15):3533-59.

35. Rabanser S, Günnemann S, Lipton Z. Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems. 2019;32.

36. Davis SE, Greevy Jr RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. Journal of biomedical informatics. 2020;112:103611.

37. Kouw WM, Loog M. An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:181211806. 2018.

38. Davis SE, Greevy Jr RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. Journal of the American Medical Informatics Association. 2019;26(12):1448-57.

39. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. A comprehensive survey on transfer learning. Proceedings of the IEEE. 2020;109(1):43-76.

40. Siregar S, Nieboer D, Versteegh MI, Steyerberg EW, Takkenberg JJ. Methods for updating a risk prediction model for cardiac surgery: a statistical primer. Interactive cardiovascular and thoracic surgery. 2019;28(3):333-8.

41. Subbaswamy A, Adams R, Saria S, editors. Evaluating model robustness and stability to dataset shift. International Conference on Artificial Intelligence and Statistics; 2021: PMLR.

42. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-52.

43. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation. 2015;131(2):211-9.

44. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB, Collins GS. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744.

45. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019;170(1):51-8.

46. Khalifa M, Magrabi F, Gallego B. Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support. BMC medical informatics and decision making. 2019;19(1):1-17.

47. Otles E, Oh J, Li B, Bochinski M, Joo H, Ortwine J, Shenoy E, Washer L, Young VB, Rao K, editors. Mind the performance gap: examining dataset shift during prospective validation. Machine Learning for Healthcare Conference; 2021: PMLR.

48. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern recognition. 2012;45(1):521-30.

49. Wan B, Caffo B, Vedula SS. A unified framework on generalizability of clinical prediction models. Frontiers in Artificial Intelligence.90.

50. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ digital medicine. 2020;3(1):1-10.

51. Dekkers O, Elm Ev, Algra A, Romijn J, Vandenbroucke J. How to assess the external validity of therapeutic trials: a conceptual approach. International journal of epidemiology. 2010;39(1):89-94.

52. Ruggles S. Big microdata for population research. Demography. 2014;51(1):287-97.

53. Sanders J, Powers B, Grossmann C. Digital data improvement priorities for continuous learning in health and health care: workshop summary: National Academies Press; 2013.

54. Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, Kimber E, Lincoln T, Mattison JE. The HL7 clinical document architecture. Journal of the American Medical Informatics Association. 2001;8(6):552-69.

55. Heymans S, McKennirey M, Phillips J. Semantic validation of the use of SNOMED CT in HL7 clinical documents. Journal of biomedical semantics. 2011;2(1):1-16.

56. Liyanage H, Krause P, De Lusignan S. Using ontologies to improve semantic interoperability in health data. BMJ Health & Care Informatics. 2015;22(2).

57. Coustasse A, Paul III DP. Adoption of the ICD-10 standard in the United States: The time is now. The Health Care Manager. 2013;32(3):260-7.

58. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR. Observational Health Data Sciences and Informatics (OHDSI):

opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574.

59. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT professional. 2005;7(5):17-23.

60. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Obtaining data from electronic health records. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User’s Guide, 3rd Edition, Addendum 2 [Internet]: Agency for Healthcare Research and Quality (US); 2019.

61. Bowles KH, Potashnik S, Ratcliffe SJ, Rosenberg MM, Shih MN-W, Topaz MM, Holmes JH, Naylor MD. Conducting research using the electronic health record across multi-hospital systems: semantic harmonization implications for administrators. The Journal of nursing administration. 2013;43(6):355.

62. Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. Journal of the American Medical Informatics Association. 2018;25(10):1292-300.

63. Sjoding MW, Iwashyna TJ, Dimick JB, Cooke CR. Gaming hospital-level pneumonia 30-day mortality and readmission measures by legitimate changes to diagnostic coding. Critical care medicine. 2015;43(5):989.

64. Lindenauer PK, Lagu T, Shieh M-S, Pekow PS, Rothberg MB. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009. Jama. 2012;307(13):1405-13.

65. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. New England Journal of Medicine. 1985;312(25):1604-8.

66. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association. 2013;20(1):117-21.

67. Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, Kohane IS. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ digital medicine. 2021;4(1):1-6.

68. Lipton ZC, Kale D, Wetzel R, editors. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. Machine learning for healthcare conference; 2016: PMLR.

69. Fu L-H, Knaplund C, Cato K, Perotte A, Kang M-J, Dykes PC, Albers D, Collins Rossetti S. Utilizing timestamps of longitudinal electronic health record data to classify clinical deterioration events. Journal of the American Medical Informatics Association. 2021;28(9):1955-63.

70. Sisk R, Lin L, Sperrin M, Barrett JK, Tom B, Diaz-Ordaz K, Peek N, Martin GP. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. Journal of the American Medical Informatics Association. 2021;28(1):155-66.

71. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. Bmj. 2018;361.

72. Sauer CM, Dam TA, Celi LA, Faltys M, de la Hoz MA, Adhikari L, Ziesemer KA, Girbes A, Thoral PJ, Elbers P. Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists. Critical care medicine. 2022.

73. Futoma J, Simons M, Doshi-Velez F, Kamaleswaran R. Generalization in Clinical Prediction Models: The Blessing and Curse of Measurement Indicator Variables. Critical Care Explorations. 2021;3(7).

74. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.

75. Chaibub Neto E, Pratap A, Perumal TM, Tummalacherla M, Snyder P, Bot BM, Trister AD, Friend SH, Mangravite L, Omberg L. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. NPJ digital medicine. 2019;2(1):1-6.

76. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation. 1998;10(7):1895-923.

77. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):56-67.

78. Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, Hinton GE. Neural additive models: Interpretable machine learning with neural nets. Advances in Neural Information Processing Systems. 2021;34.

79. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. Journal of biomedical informatics. 2014;51:24-34.

80. Zheng K, Gao J, Ngiam KY, Ooi BC, Yip WLJ, editors. Resolving the bias in electronic medical records. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017.

81. Fleming SL, Jeyapragasan K, Duan T, Ding D, Gombar S, Shah N, Brunskill E. Missingness as stability: Understanding the structure of missingness in longitudinal ehr data and its impact on reinforcement learning in healthcare. arXiv preprint arXiv:191107084. 2019.

82. Shelly MP, Nightingale P. Respiratory support. Bmj. 1999;318(7199):1674-7.

83. Vincent J-L, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C, Suter P, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Springer-Verlag; 1996.

84. Yadaw AS, Li Y-c, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. The Lancet Digital Health. 2020;2(10):e516-e25.

85. Chute CG. Clinical classification and terminology: some history and current observations. Journal of the American Medical Informatics Association. 2000;7(3):298-303.

86. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical chemistry. 2003;49(4):624-33.

87. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl_1):D267-D70.

88. Nishimura M. High-flow nasal cannula oxygen therapy in adults. Journal of intensive care. 2015;3(1):1-8.

89. Xu Q, Wang T, Qin X, Jie Y, Zha L, Lu W. Early awake prone position combined with high-flow nasal oxygen therapy in severe COVID-19: a case series. Critical Care. 2020;24(1):1-3.

90. Ang K, Green A, Ramaswamy K, Frerk C. Preoxygenation using the Optiflow™ system. BJA: British Journal of Anaesthesia. 2017;118(3):463-4.

91. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics. 2006;121:279.

92. Cao H, Lee K, Ennett CM, Eshelman L, Nielsen L, Saeed M, Gross B, editors. Heuristics to determine ventilation times of ICU patients from the MIMIC-II database. 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology; 2010: IEEE.

93. Roy S, Mincu D, Loreaux E, Mottram A, Protsyuk I, Harris N, Xue Y, Schrouff J, Montgomery H, Connell A. Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. Journal of the American Medical Informatics Association. 2021.

94. Mody A, Lyons PG, Vazquez Guillamet C, Michelson A, Yu S, Namwase AS, Sinha P, Powderly WG, Woeltje K, Geng EH. The Clinical Course of Coronavirus Disease 2019 in a US Hospital System: A Multistate Analysis. American Journal of Epidemiology. 2021;190(4):539-52.

95. Demšar J. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research. 2006;7:1-30.

96. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems; 2017.

97. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17(3):261-72.

98. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ. Array programming with NumPy. Nature. 2020;585(7825):357-62.

99. McKinney W, editor Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference; 2010: Austin, TX.

100. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007;9(03):90-5.

101. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

102. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.

103. Wang Y, Lu X, Li Y, Chen H, Chen T, Su N, Huang F, Zhou J, Zhang B, Yan F. Clinical course and outcomes of 344 intensive care patients with COVID-19. American journal of respiratory and critical care medicine. 2020;201(11):1430-4.

104. Bobroske K, Larish C, Cattrell A, Bjarnadóttir MV, Huan L. The bird's-eye view: A data-driven approach to understanding patient journeys from claims data. Journal of the American Medical Informatics Association. 2020;27(7):1037-45.

105. Patel BK, Wolfe KS, Pohlman AS, Hall JB, Kress JP. Effect of noninvasive ventilation delivered by helmet vs face mask on the rate of endotracheal intubation in patients with acute respiratory distress syndrome: a randomized clinical trial. Jama. 2016;315(22):2435-41.

106. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical care. 1998:8-27.

107. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J-C, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care. 2005:1130-9.

108. Moore BJ, White S, Washington R, Coenen N, Elixhauser A. Identifying increased risk of readmission and in-hospital mortality using hospital administrative data. Medical care. 2017;55(7):698-705.

109. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical care. 2009:626-33.

110. Sharma N, Schwendimann R, Endrich O, Ausserhofer D, Simon M. Comparing Charlson and Elixhauser comorbidity indices with different weightings to predict in-hospital mortality: an analysis of national inpatient data. BMC health services research. 2021;21(1):1-10.

111. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artificial intelligence in medicine. 2016;71:57-61.

112. Angus DC, Seymour CW, Coopersmith CM, Deutschman C, Klompas M, Levy MM, Martin GS, Osborn TM, Rhee C, Watson RS. A framework for the development and interpretation of different sepsis definitions and clinical criteria. Crit Care Med. 2016;44(3):e113.

113. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, Hammond W, Califf RM, Spratt SE. A comparison of phenotype definitions for diabetes mellitus. Journal of the American Medical Informatics Association. 2013;20(e2):e319-e26.

114. Glynn EF, Hoffman MA. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. JAMIA open. 2019;2(4):554-61.

115. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2014;21(2):221-30.

116. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. Journal of biomedical informatics. 2015;58:156-65.

117. Buchman TG, Simpson SQ, Sciarretta KL, Finne KP, Sowers N, Collier M, Chavan S, Oke I, Pennini ME, Santhosh A. Sepsis among medicare beneficiaries: 1. The burdens of sepsis, 2012–2018. Critical care medicine. 2020;48(3):276.

118. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016;315(8):801-10.

119. Torio C, Moore BJ. National inpatient hospital costs: the most expensive conditions by payer, 2013: statistical brief #204. 2006.

120. Saria S, Henry KE. Too Many Definitions of Sepsis: Can Machine Learning Leverage the Electronic Health Record to Increase Accuracy and Bring Consensus? Critical Care Medicine. 2020;48(2):137-41.

121. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, Schein RM, Sibbald WJ. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest. 1992;101(6):1644-55.

122. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent J-L, Ramsay G. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. Intensive Care Med. 2003;29(4):530-8.

123. Vincent J-L. Dear SIRS, I'm sorry to say that I don't like you. Crit Care Med. 1997;25(2):372-4.

124. Churpek MM, Zadravecz FJ, Winslow C, Howell MD, Edelson DP. Incidence and prognostic value of the systemic inflammatory response syndrome and organ dysfunctions in ward patients. Am J Respir Crit Care Med. 2015;192(8):958-64.

125. Kaukonen K-M, Bailey M, Pilcher D, Cooper DJ, Bellomo R. Systemic inflammatory response syndrome criteria in defining severe sepsis. N Engl J Med. 2015;372(17):1629-38.

126. Specifications manual for national hospital inpatient quality measures. The Joint Commission; 2014.

127. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M. Assessment of clinical criteria for sepsis: for the Third

International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-74.

128. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, Kadri SS, Angus DC, Danner RL, Fiore AE. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. JAMA. 2017;318(13):1241-9.

129. Simpson SQ. New sepsis criteria: a change we should not make. Chest. 2016;149(5):1117-8.

130. Simpson SQ. SIRS in the time of Sepsis-3. Chest. 2018;153(1):34-8.

131. Townsend SR, Rivers E, Tefera L. Definitions for sepsis and septic shock. JAMA. 2016;316(4):457-8.

132. Fang X, Wang Z, Yang J, Cai H, Yao Z, Li K, Fang Q. Clinical evaluation of Sepsis-1 and Sepsis-3 in the ICU. Chest. 2018;153(5):1169-76.

133. Cheng B, Li Z, Wang J, Xie G, Liu X, Xu Z, Chu L, Zhao J, Yao Y, Fang X. Comparison of the performance between sepsis-1 and sepsis-3 in ICUs in China: a retrospective multicenter study. Shock. 2017;48(3):301.

134. Szakmany T, Pugh R, Kopczynska M, Lundin RM, Sharif B, Morgan P, Ellis G, Abreu J, Kulikouskaya S, Bashir K. Defining sepsis on the wards: results of a multi-centre point-prevalence study comparing two sepsis definitions. Anaesthesia. 2018;73(2):195-204.

135. Poutsiaka DD, Porto MC, Perry WA, Hudcova J, Tybor DJ, Hadley S, Doron S, Reich JA, Snydman DR, Nasraway SA. Prospective observational study comparing sepsis-2 and sepsis-3 definitions in predicting mortality in critically ill patients. Open Forum Infectious Diseases. 2019;6(7):ofz271.

136. Serafim R, Gomes JA, Salluh J, Póvoa P. A comparison of the quick-SOFA and systemic inflammatory response syndrome criteria for the diagnosis of sepsis and prediction of mortality: a systematic review and meta-analysis. Chest. 2018;153(3):646-55.

137. Gando S, Shiraishi A, Abe T, Kushimoto S, Mayumi T, Fujishima S, Hagiwara A, Shiino Y, Shiraishi S-i, Hifumi T. The SIRS criteria have better performance for predicting infection than qSOFA scores in the emergency department. Sci Rep. 2020;10(1):1-9.

138. Johnson AE, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Celi LA, Stone DJ. A comparative analysis of sepsis identification methods in an electronic database. Crit Care Med. 2018;46(4):494.

139. Henry KE, Hager DN, Osborn TM, Wu AW, Saria S. Comparison of Automated Sepsis Identification Methods and Electronic Health Record–based Sepsis Phenotyping: Improving Case Identification Accuracy by Accounting for Confounding Comorbid Conditions. Critical Care Explorations. 2019;1(10):e0053.

140. Albritton N, Raveendran R, Jackson C. SEP-1 Early Management Bundle, Severe Sepsis/Septic Shock: V5.4 Measure Updates. Quesitons and answers transcript.: Joint Comission; 2018.

141. Bauer SR, Gonet JA, Rosario RF, Griffiths LA, Kingery T, Reddy AJ. Inter-rater Agreement for Abstraction of the Early Management Bundle, Severe Sepsis/Septic Shock (SEP-1) Quality Measure in a Multi-Hospital Health System. The Joint Commission Journal on Quality and Patient Safety. 2019;45(2):108-11.

142. Rhee C, Filbin M, Massaro AF, Bulger A, McEachern D, Tobin KA, Kitch B, Thurlo-Walsh B, Kadar A, Koffman A. Compliance with the national SEP-1 quality measure and association with sepsis outcomes: a multicenter retrospective cohort study. Crit Care Med. 2018;46(10):1585.

143. Pepper DJ, Sun J, Cui X, Welsh J, Natanson C, Eichacker PQ. Antibiotic-and fluid-focused bundles potentially improve sepsis management, but high-quality evidence is lacking for the specificity required in the Centers for Medicare and Medicaid Service's sepsis bundle (SEP-1). Crit Care Med. 2019;47(10):1290-300.

144. Mackay F, Roy A, Schorr C, Crabtree P, Puri N. 1471: CMS SEP-1 MEASURE START TIME: DO WE AGREE? A COMPARISON OF CLINICIANS VERSUS QUALITY STAFF. Crit Care Med. 2018;46(1):719.

145. Song X, Yu AS, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, Hu Y, Liu M. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nature communications. 2020;11(1):1-12.

146. Raschka S. An overview of general performance metrics of binary classifier systems. arXiv preprint arXiv:14105330. 2014.

147. Kuhn M, Vaughan D. yardstick: Tidy Characterizations of Model Performance. R package version 0.0. 6. 2020.

148. Lauritsen SM, Thiesson B, Jørgensen MJ, Riis AH, Espelund US, Weile JB, Lange J. The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. NPJ digital medicine. 2021;4(1):1-12.

149. Lee TC, Shah NU, Haack A, Baxter SL, editors. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics; 2020: Multidisciplinary Digital Publishing Institute.

150. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, Bock C, Horn M, Moor M, Rieck B. Early prediction of circulatory failure in the intensive care unit using machine learning. Nature medicine. 2020;26(3):364-73.

151. Marwaha JS, Kvedar JC. Crossing the chasm from model performance to clinical impact: the need to improve implementation and evaluation of AI. Nature Publishing Group; 2022. p. 1-2.

152. Lee C, Lawson BL, Mann AJ, Liu VX, Myers LC, Schuler A, Escobar GJ. Exploratory analysis of novel electronic health record variables for quantification of healthcare delivery strain, prediction of mortality, and prediction of imminent discharge. Journal of the American Medical Informatics Association. 2022.

153. Karakra A, Fontanili F, Lamine E, Lamothe J. A discrete event simulation-based methodology for building a digital twin of patient pathways in the hospital for near real-time monitoring and predictive simulation. Digital Twin. 2022;2(1):1.

154. Williams E, Szakmany T, Spernaes I, Muthuswamy B, Holborn P. Discrete-event simulation modeling of critical care flow: New hospital, old challenges. Critical care explorations. 2020;2(9).

155. Zhang X. Application of discrete event simulation in health care: a systematic review. BMC health services research. 2018;18(1):1-11.

156. DeRienzo CM, Shaw RJ, Meanor P, Lada E, Ferranti J, Tanaka D. A discrete event simulation tool to support and predict hospital and clinic staffing. Health informatics journal. 2017;23(2):124-33.

157. Dittus RS. Discrete-event simulation modeling of the content, processes, and structures of health care: Building a Better Delivery System: A New Engineering/Health Care Partnership …; 2005.

158. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Mottram A, Meyer C, Ravuri S, Protsyuk I. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116-9.

159. Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE–an interpretable and recurrent neural survival model for early prediction of sepsis. Artificial Intelligence in Medicine. 2021;113:102036.

160. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine. 2018;1(1):1-10.

161. Mišić VV, Rajaram K, Gabel E. A simulation-based evaluation of machine learning models for clinical decision support: application and analysis using hospital readmission. NPJ Digital Medicine. 2021;4(1):1-11.

162. Shah PK, Ginestra JC, Ungar LH, Junker P, Rohrbach JI, Fishman NO, Weissman GE. A Simulated Prospective Evaluation of a Deep Learning Model for Real-Time Prediction of Clinical Deterioration Among Ward Patients. Critical Care Medicine. 2021.

163. Liang L, Moore B, Soni A. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017: Statistical Brief# 261. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]. 2006.

164. Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, Iwashyna TJ. Hospital deaths in patients with sepsis from 2 independent cohorts. JAMA. 2014;312(1):90-2.

165. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine. 2017;376(23):2235-44.

166. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine. 2006;34(6):1589-96.

167. Pimentel MA, Redfern OC, Gerry S, Collins GS, Malycha J, Prytherch D, Schmidt PE, Smith GB, Watkinson PJ. A comparison of the ability of the National Early Warning Score and the

National Early Warning Score 2 to identify patients at risk of in-hospital mortality: a multi-centre database study. Resuscitation. 2019;134:147-56.

168. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Critical Care Medicine. 2018;46(4):547.

169. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Medical Informatics. 2016;4(3):e5909.

170. Reyna MA, Josef C, Seyedi S, Jeter R, Shashikumar SP, Westover MB, Sharma A, Nemati S, Clifford GD, editors. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. 2019 Computing in Cardiology (CinC); 2019: IEEE.

171. Levy MM, Dellinger RP, Townsend SR, Linde-Zwirble WT, Marshall JC, Bion J, Schorr C, Artigas A, Ramsay G, Beale R. The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. Intensive Care Medicine. 2010;36(2):222-31.

172. Yu S, Betthauser KD, Gupta A, Lyons PG, Lai AM, Kollef MH, Payne PR, Michelson AP. Comparison of Sepsis Definitions as Automated Criteria. Critical Care Medicine. 2021.

173. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. Xgboost: extreme gradient boosting. R package version 04-2. 2015;1(4).

174. Yu SC, Shivakumar N, Betthauser K, Gupta A, Lai AM, Kollef MH, Payne PR, Michelson AP. Performance of Early Warning Scores for Sepsis Identification in the General Ward Setting. Journal of the American Medical Informatics Association Open. In Press.

175. Hunter JD. Matplotlib: A 2D graphics environment. IEEE Annals of the History of Computing. 2007;9(03):90-5.

176. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

177. Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, Swart EL, Girbes AR, Thoral P, Ercole A. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Medicine. 2020;46(3):383-400.

178. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, Kalet IJ, Lenert LA, Musen MA, Ozbolt JG. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. Journal of the American Medical Informatics Association. 2012;19(6):931-8.

179. Craven CK, Doebbeling B, Furniss D, Holden RJ, Lau F, Novak LL. Evidence-based health informatics frameworks for applied use. Stud Health Technol Inform. 2016;222:77-89.

180. Nilsen P. Making sense of implementation theories, models, and frameworks: Implementation science. Springer: Philadelphia, PA, USA; 2015.

181. Ladan MA, Wharrad H, Windle R. Towards understanding healthcare professionals' adoption and use of technologies in clinical practice: using Q-methodology and models of technology acceptance. BMJ Health & Care Informatics. 2018;25(1).

182. Holahan PJ, Lesselroth BJ, Adams K, Wang K, Church V. Beyond technology acceptance to effective technology use: a parsimonious and actionable model. Journal of the American Medical Informatics Association. 2015;22(3):718-29.

183. Sittig DF, Singh H. A new socio-technical model for studying health information technology in complex adaptive healthcare systems. Cognitive informatics for biomedicine: Springer; 2015. p. 59-80.

184. Ubel PA, Rosenthal MB. Beyond nudges—when improving health calls for greater assertiveness. New England Journal of Medicine. 2019;380(4):309-11.

185. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior–a new perspective on a persistent problem. Journal of the American Medical Informatics Association. 2017;24(2):409-12.

186. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR medical informatics. 2018;6(2):e8912.

187. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. Journal of medical Internet research. 2020;22(6):e15154.

188. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. Jama. 2019;322(18):1765-6.

189. Barda N, Yona G, Rothblum GN, Greenland P, Leibowitz M, Balicer R, Bachmat E, Dagan N. Addressing bias in prediction models by improving subpopulation calibration. Journal of the American Medical Informatics Association. 2021;28(3):549-58.

190. Aaron S, McEvoy DS, Ray S, Hickman T-TT, Wright A. Cranky comments: detecting clinical decision support malfunctions through free-text override reasons. Journal of the American Medical Informatics Association. 2019;26(1):37-43.

191. Yu SC, Hofford MR, Lai AM, Kollef MH, Payne PR, Michelson AP. Respiratory support status from EHR data for adult population: classification, heuristics, and usage in predictive modeling. Journal of the American Medical Informatics Association. 2022.

192. Yu SC, Betthauser KD, Gupta A, Lyons PG, Lai AM, Kollef MH, Payne PR, Michelson AP. Comparison of sepsis definitions as automated criteria. Critical care medicine. 2021;49(4):e433-e43.

193. Yu SC, Gupta A, Betthauser K, Lyons PG, Lai AM, Kollef MH, Payne PR, Michelson AP. Sepsis Prediction for the General Ward Setting. Frontiers in Digital Health. 2022:21.

194. McDonald LC, Gerding DN, Johnson S, Bakken JS, Carroll KC, Coffin SE, Dubberke ER, Garey KW, Gould CV, Kelly C. Clinical practice guidelines for Clostridium difficile infection in adults and children: 2017 update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). Clinical infectious diseases. 2018;66(7):e1-e48.

Appendices

Appendix 1. TRIPOD

Section/Topic	Item	Checklist Item
Title and abstract		
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Introduction		
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.
Methods		
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.
	5b	Describe eligibility criteria for participants.
	5c	Give details of treatments received, if relevant.
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	Explain how the study size was arrived at.
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	For validation, describe how the predictions were calculated.
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	Provide details on how risk groups were created, if done.
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Results		
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).
Model development	14a	Specify the number of participants and outcome events in each analysis.
	14b	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	Explain how to the use the prediction model.
Model performance	16	Report performance measures (with CIs) for the prediction model.
Model-updating	17	If done, report the results from any model updating (i.e., model specification, model performance).
Discussion		
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	Discuss the potential clinical use of the model and implications for future research.
Other information		
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.
Funding	22	Give the source of funding and the role of the funders for the present study.

Appendix 2. CHARMS

Domain	#	Key items
SOURCE OF DATA	1	Source of data (e.g., cohort, case-control, randomized trial participants, or registry data)
PARTICIPANTS	2	Participant eligibility and recruitment method (e.g., consecutive participants, location, number of centers, setting, inclusion and exclusion criteria)
	3	Participant description
	4	Details of treatments received, if relevant
	5	Study dates
OUTCOME(S) TO BE PREDICTED	6	Definition and method for measurement of outcome
	7	Was the same outcome definition (and method for measurement) used in all patients?
	8	Type of outcome (e.g., single or combined endpoints)
	9	Was the outcome assessed without knowledge of the candidate predictors (i.e., blinded)?
	10	Were candidate predictors part of the outcome (e.g., in panel or consensus diagnosis)?
	11	Time of outcome occurrence or summary of duration of follow-up
CANDIDATE PREDICTORS (OR INDEX TESTS)	12	Number and type of predictors (e.g., demographics, patient history, physical examination, additional testing, disease characteristics)
	13	Definition and method for measurement of candidate predictors
	14	Timing of predictor measurement (e.g., at patient presentation, at diagnosis, at treatment initiation)
	15	Were predictors assessed blinded for outcome, and for each other (if relevant)?
	16	Handling of predictors in the modelling (e.g., continuous, linear, non-linear transformations or categorised)
SAMPLE SIZE	17	Number of participants and number of outcomes/events
	18	Number of outcomes/events in relation to the number of candidate predictors (Events Per Variable)
MISSING DATA	19	Number of participants with any missing value (include predictors and outcomes)
	20	Number of participants with missing data for each predictor
	21	Handling of missing data (e.g., complete-case analysis, imputation, or other methods)
MODEL DEVELOPMENT	22	Modelling method (e.g., logistic, survival, neural network, or machine learning techniques)
	23	Modelling assumptions satisfied
	24	Method for selection of predictors **for inclusion** in multivariable modelling (e.g., all candidate predictors, pre-selection based on unadjusted association with the outcome)
	25	Method for selection of predictors **during multivariable modelling** (e.g., full model approach, backward or forward selection) and criteria used (e.g., p-value, Akaike Information Criterion)
	26	Shrinkage of predictor weights or regression coefficients (e.g., no shrinkage, uniform shrinkage, penalized estimation)
MODEL PERFORMANCE	27	Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination
	28	(C-statistic, D-statistic, log-rank) measures with confidence intervals
	29	Classification measures (e.g., sensitivity, specificity, predictive values, net reclassification improvement) and whether a-priori cut points were used
MODEL	30	Method used for testing model performance: development dataset only (random split of data, resampling methods e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators)
EVALUATION	31	In case of poor validation, whether model was adjusted or updated (e.g., intercept recalibrated, predictor effects adjusted, or new predictors added)
RESULTS	32	Final and other multivariable models (e.g., basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals)
	33	Any alternative presentation of the final prediction models, e.g., sum score, nomogram, score chart, predictions for specific risk subgroups with performance
	34	Comparison of the distribution of predictors (including missing data) for development and validation datasets
INTERPRETATION AND DISCUSSION	35	Interpretation of presented models (confirmatory, i.e., model useful for practice versus exploratory, i.e., more research needed)
	36	Comparison with other studies, discussion of generalizability, strengths and limitations.

Appendix 3. PROBAST

Domain	#	Item
Participants	**1.1**	Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?
	1.2	Were all inclusions and exclusions of participants appropriate?
Predictors	**2.1**	Were predictors defined and assessed in a similar way for all participants?
	2.2	Were predictor assessments made without knowledge of outcome data?
	2.3	Are all predictors available at the time the model is intended to be used?
Outcome	**3.1**	Was the outcome determined appropriately?
	3.2	Was a pre-specified or standard outcome definition used?
	3.3	Were predictors excluded from the outcome definition?
	3.4	Was the outcome defined and determined in a similar way for all participants?
	3.5	Was the outcome determined without knowledge of predictor information?
	3.6	Was the time interval between predictor assessment and outcome determination appropriate?
Analysis	**4.1**	Were there a reasonable number of participants with the outcome?
	4.2	Were continuous and categorical predictors handled appropriately?
	4.3	Were all enrolled participants included in the analysis?
	4.4	Were participants with missing data handled appropriately?
	4.5	Was selection of predictors based on univariable analysis avoided?
	4.6	Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?
	4.7	Were relevant model performance measures evaluated appropriately?
	4.8	Were model overfitting and optimism in model performance accounted for?
	4.9	Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?

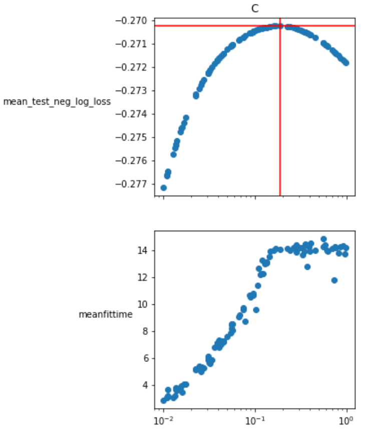
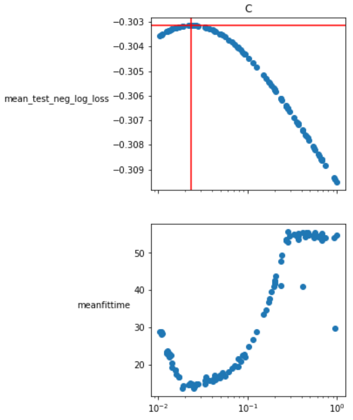
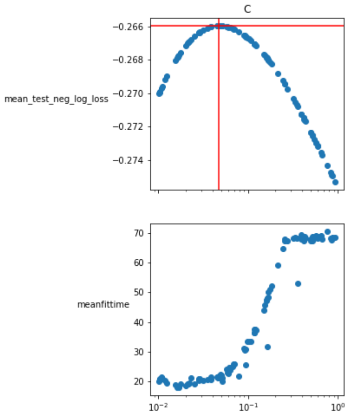
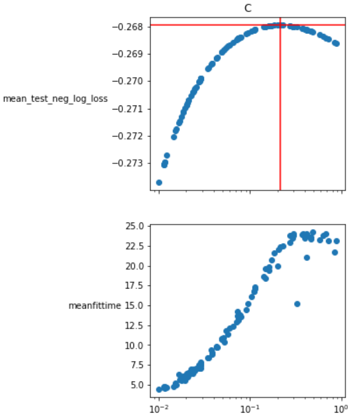
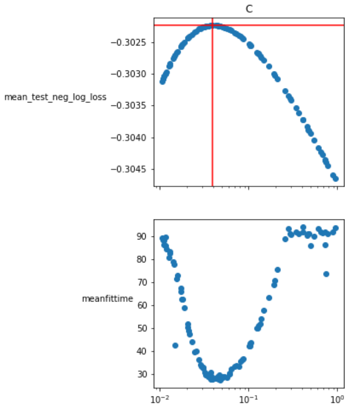
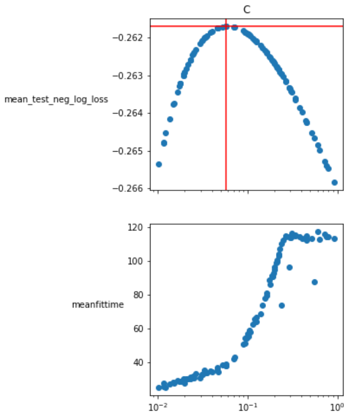
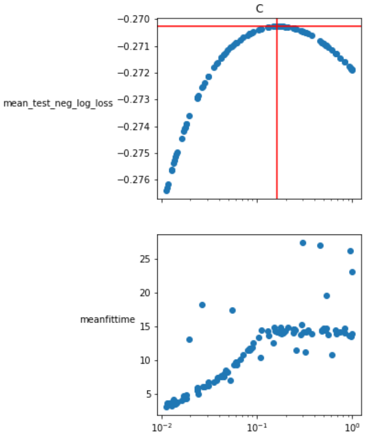
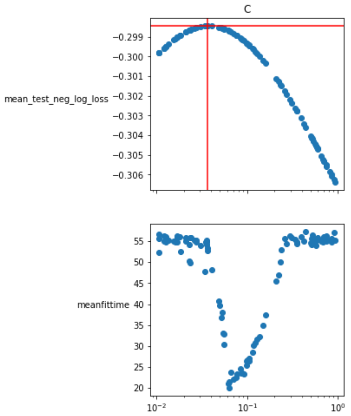
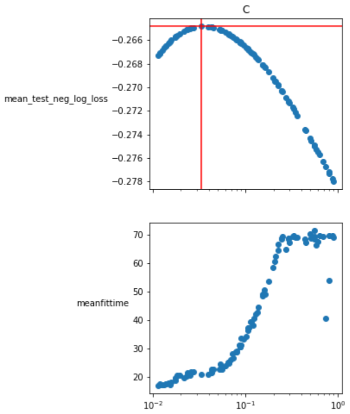
Appendix 4. GRASP

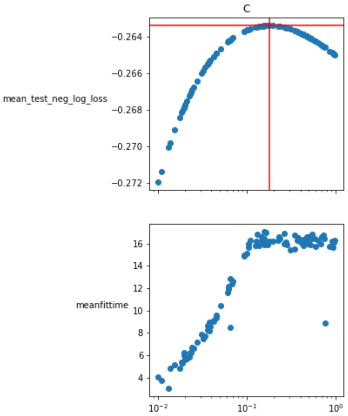
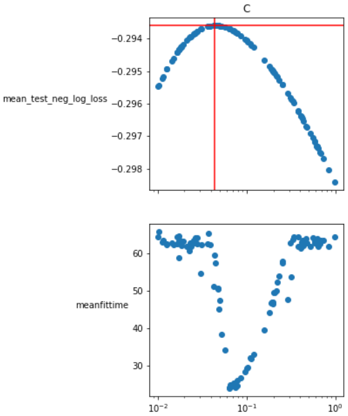
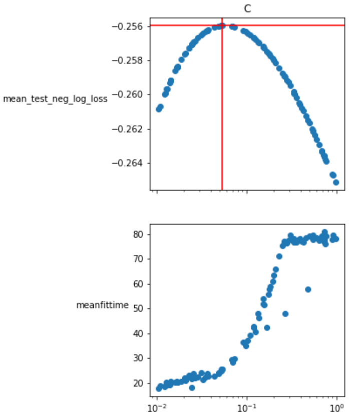
Category	Item	#	Description
Background	**Name**	**1**	Name of predictive tool (report tool's creators and year in the absence of a given name)
	Author	**2**	Name of developer (first author or researcher)
	Country	**3**	Country of development
	Year	**4**	Year of development
	Category	**5**	Diagnostic/Therapeutic/Prognostic/Preventive
	Intended use	**6**	Specific aim/intended use of the predictive tool
	Intended user	**7**	Type of practitioner intended to use the tool
	Clinical area	**8**	Clinical specialty
	Target population	**9**	Target patient population and health care settings in which the tool is applied
	Taregt outcome	**10**	Event to be predicted (including prediction lead time if needed)
	Action	**11**	Recommended action based on tool's output
	Input source	**12**	Clinical (including Diagnostic, Genetic, Vital signs, Pathology) or non-clinical (including Healthcare Utilisation)
	Input type	**13**	Objective (Measured input; from electronic systems or clinical examination) or subjective (Patient reported; history, checklist …etc.)
	Local context	**14**	Is the tool developed using location-specific data? (e.g. life expectancy tables)
	Methodology	**15**	Type of algorithm used for developing the tool (e.g. parametric/non-parametric)
	Internal validation	**16**	Method of internal validation
	Dedicated support	**17**	Name of the supporting/funding research networks, programs, or professional groups
	Endorsement	**18**	Organisations endorsing the tool and/or clinical guidelines recommending its utilisation
	Automation flag	**19**	Automation status (manual/automated)
	Tool citations	**20**	Total citations of the tool
	Studies	**21**	Number of studies reporting the tool
	Author #	**22**	Number of authors
	Sample size	**23**	Size of patient/record sample used in the development of the tool
	Journal Name	**24**	Name of the journal that published the tool's primary development study
	Journal Rank	**25**	Impact factor of the journal
	Citation Index	**26**	Calculated as: Average Annual Citations = number of citations/age of primary publication
	Publication Index	**27**	Calculated as: Average Annual Studies = number of studies/age of primary publication
	Literature Index	**28**	Calculated as: Citations and Publications = number of citations X number of studies
Pre-implementation	**Internal validation**	**29**	Tested for internal validity
	External validation	**30**	Tested for external validity
During implementation	**Usability**	**31**	Reported usability testing
	Potential effect	**32**	Reported estimated potential effect on clinical effectiveness, patient safety, or healthcare efficiency
Post-implementation	**Subjective/descriptive**	**33**	Based on subjective studies; e.g. the opinion of a respected authority, clinical experience, a descriptive study, or a report of an expert committee or panel
	Observational	**34**	Based on observational studies; e.g. a well-designed cohort or case-control study
	Experimental (impact)	**35**	Based on experimental studies; e.g. a well-designed, widely applied randomized/nonrandomized controlled trial

Appendix 5. Fixed and Optimized Parameters

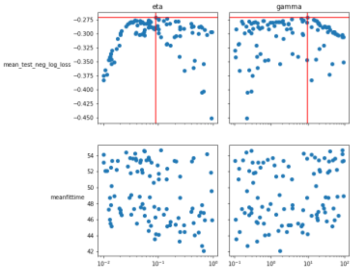
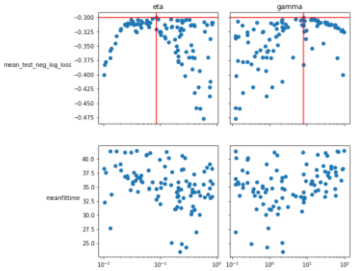
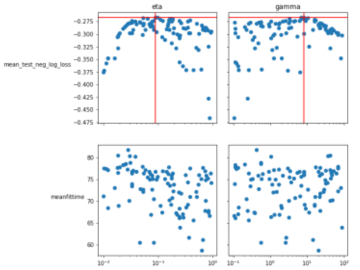
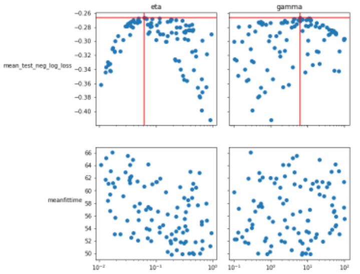
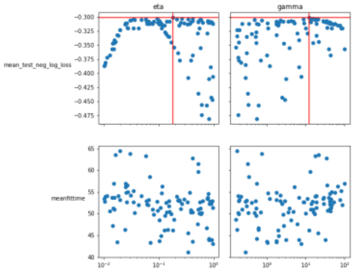
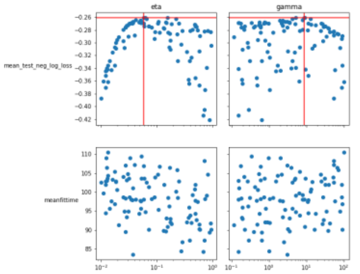
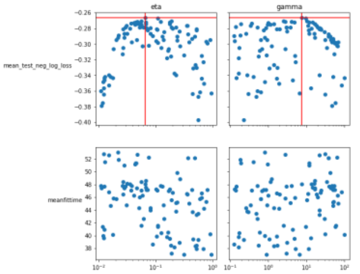
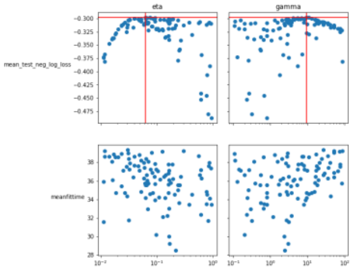
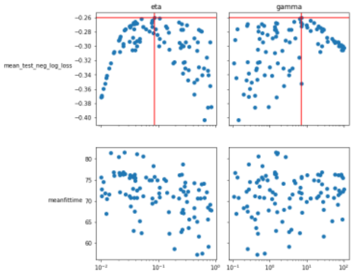
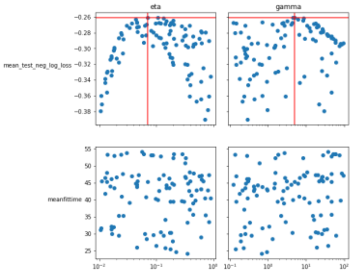
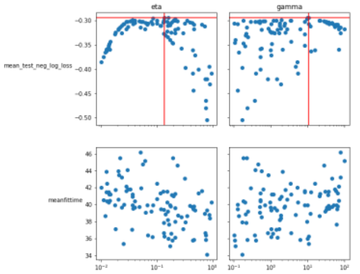
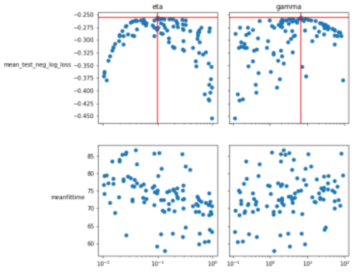
Model Type	**Fixed Parameters**	**Optimized Parameters**
LR	N/A	N/A
SplineLR opt	• (Spline) Number of knots: 5 • (Spline) Position of knots: quantiles • (Spline) Polynomial degree: 3 • Penalty type: L2	• Regularization strength, C: log-uniform between 1e-2 to 1e0
XGB opt	• Tree method: histogram-based • Number of estimators: 100 • Max tree depth: 10	• Learning rate, eta: log-uniform between 1e-2 to 1e0 • Regularization strength, gamma: log-uniform between 1e-1 to 1e2

Appendix 6. Hyperparameter Optimization Results

SplineLR	PP	HCP	HCP+PP
BJH 0			
BJH 1			
MIMIC 0			

MIMIC 1			

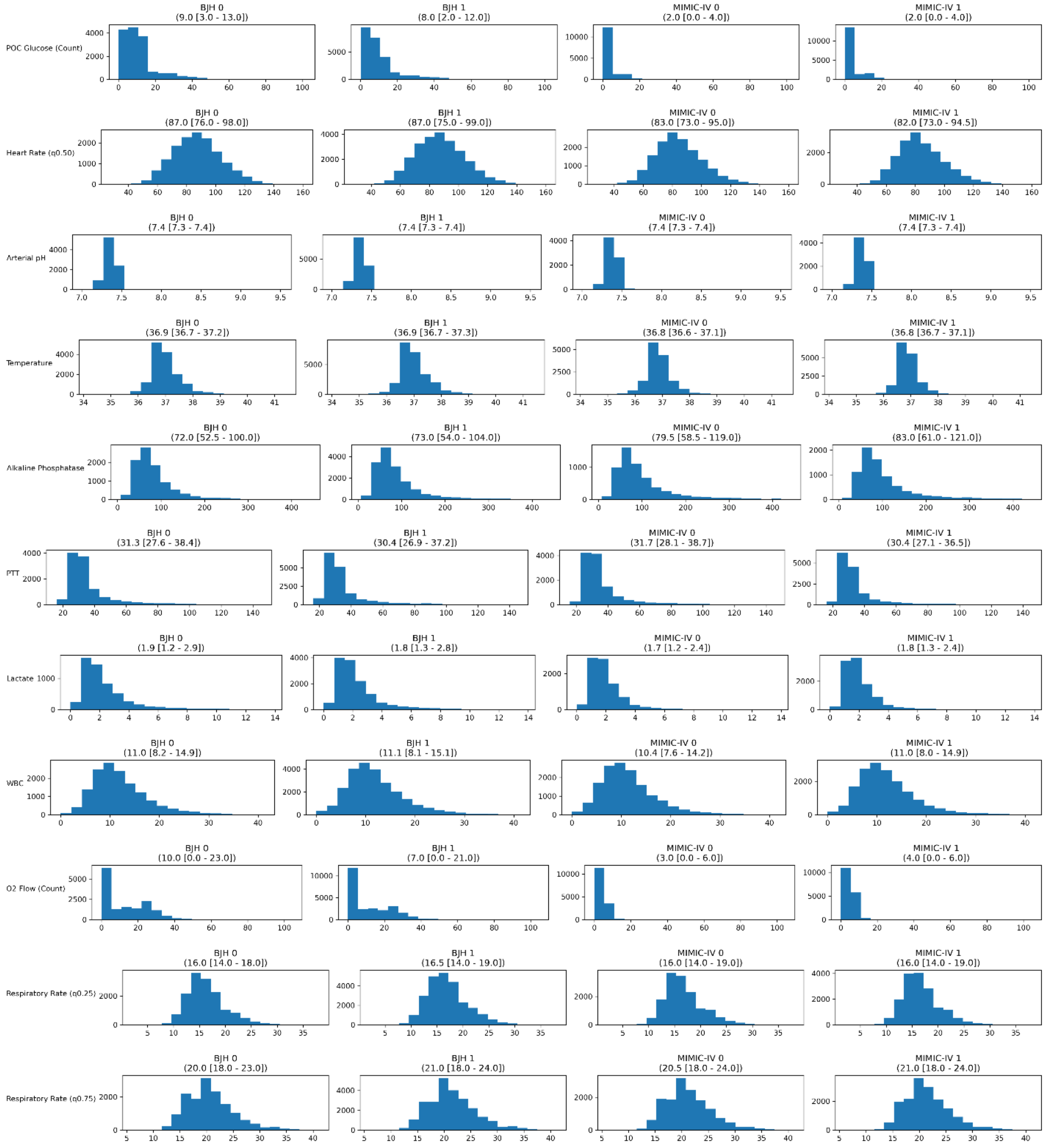
The regularization strength parameter of logistic regression, named C in the sklearn package’s implementation, was optimized through 100 iterations of 5-fold cross validation for each site-era combination. The rest of the parameters including those for spline transformation (e.g. number of knots, placement of knots, etc.) were fixed.

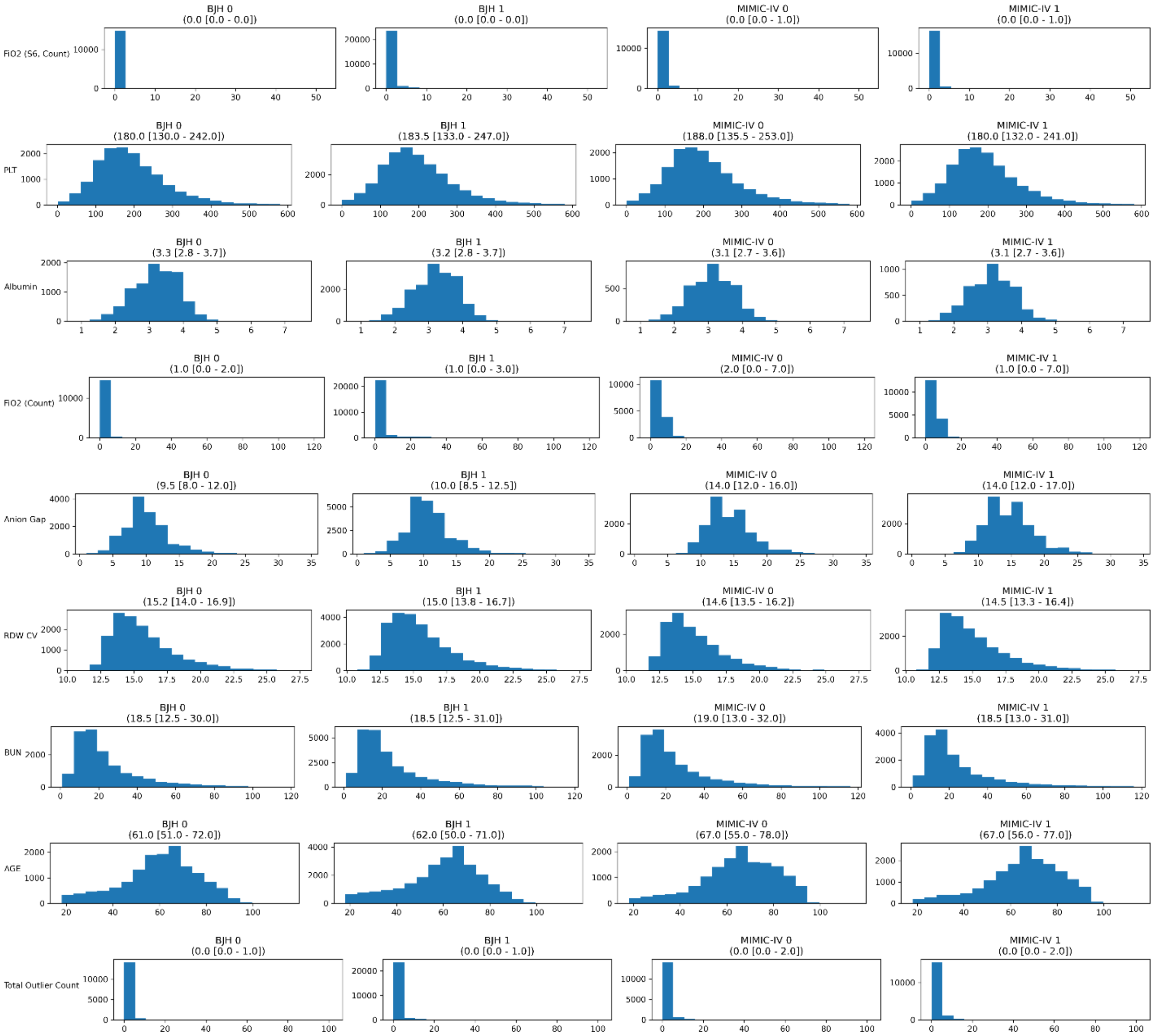
XGB	PP	HCP	HCP+PP
BJH 0			
BJH 1			
MIMIC 0			
MIMIC 1			

The learning rate and regularization strength parameter of XGBoost, named eta nd gamma respectively, were optimized through 100 iterations of 5-fold cross validation for each site-era combination. The rest of the parameters (e.g. number of trees, method of tree growth, etc.) were fixed.

Appendix 7. Comparison of Important HCP and PP Features for XGB opt

For each of the top 20 important features determined using median SHAP values across different site-era combinations for XGB opt using both HCP and PP features, the distribution histogram was generated and computed as median and IQR.




FiO2 (S6, Count)
BJH 0 (0.0 [0.0 - 0.0])
BJH 1 (0.0 [0.0 - 0.0])
MIMIC-IV 0 (0.0 [0.0 - 1.0])
MIMIC-IV 1 (0.0 [0.0 - 1.0])
0 10 20 30 40 50
PLT
BJH 0 (180.0 [130.0 - 242.0])
BJH 1 (183.5 [133.0 - 247.0])
MIMIC-IV 0 (188.0 [135.5 - 253.0])
MIMIC-IV 1 (180.0 [132.0 - 241.0])
0 100 200 300 400 500 600
Albumin
BJH 0 (3.3 [2.8 - 3.7])
BJH 1 (3.2 [2.8 - 3.7])
MIMIC-IV 0 (3.1 [2.7 - 3.6])
MIMIC-IV 1 (3.1 [2.7 - 3.6])
1 2 3 4 5 6 7
FiO2 (Count)
BJH 0 (1.0 [0.0 - 2.0])
BJH 1 (1.0 [0.0 - 3.0])
MIMIC-IV 0 (2.0 [0.0 - 7.0])
MIMIC-IV 1 (1.0 [0.0 - 7.0])
0 20 40 60 80 100 120
Anion Gap
BJH 0 (9.5 [8.0 - 12.0])
BJH 1 (10.0 [8.5 - 12.5])
MIMIC-IV 0 (14.0 [12.0 - 16.0])
MIMIC-IV 1 (14.0 [12.0 - 17.0])
0 5 10 15 20 25 30 35
RDW CV
BJH 0 (15.2 [14.0 - 16.9])
BJH 1 (15.0 [13.8 - 16.7])
MIMIC-IV 0 (14.6 [13.5 - 16.2])
MIMIC-IV 1 (14.5 [13.3 - 16.4])
10.0 12.5 15.0 17.5 20.0 22.5 25.0 27.5
BUN
BJH 0 (18.5 [12.5 - 30.0])
BJH 1 (18.5 [12.5 - 31.0])
MIMIC-IV 0 (19.0 [13.0 - 32.0])
MIMIC-IV 1 (18.5 [13.0 - 31.0])
0 20 40 60 80 100 120
AGE
BJH 0 (61.0 [51.0 - 72.0])
BJH 1 (62.0 [50.0 - 71.0])
MIMIC-IV 0 (67.0 [55.0 - 78.0])
MIMIC-IV 1 (67.0 [56.0 - 77.0])
20 40 60 80 100
Total Outlier Count
BJH 0 (0.0 [0.0 - 1.0])
BJH 1 (0.0 [0.0 - 1.0])
MIMIC-IV 0 (0.0 [0.0 - 2.0])
MIMIC-IV 1 (0.0 [0.0 - 2.0])
0 20 40 60 80 100

Appendix 8. Feature Sets for In-hospital Mortality Prediction Feature Ablation Study

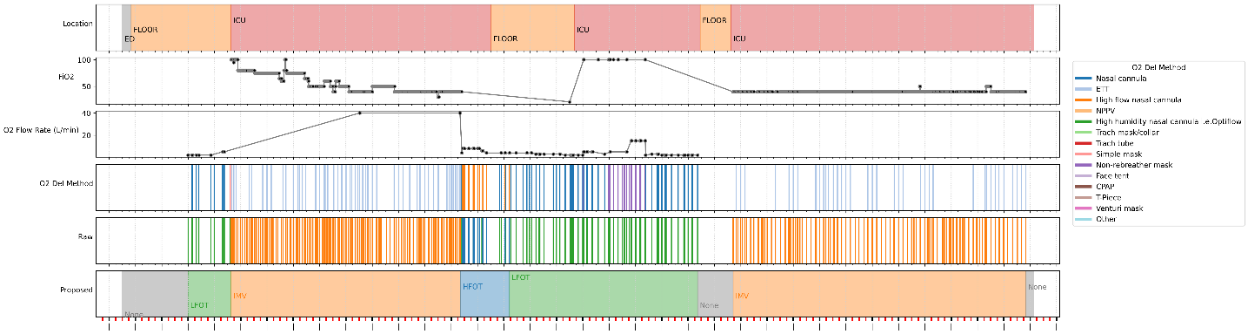
Feature set	Description	Included[a]	Features
Baseline	Features easily acquirable and commonly used for EHR-based prediction modeling from vital signs and lab results		Age, BMI, Sex (male), Race **(25th, 50th, and 75th quantiles and IQR)** Heart rate, MAP, Respiratory rate, SpO2, Temperature **(median)** BMI, GCS eye, GCS motor, GCS total, GCS verbal, Heart rate, MAP, Respiratory rate, SpO2, Temperature, ALP, ALT, ASP, Albumin, Anion gap, Basophil (absolute), Basophil (percent), Bilirubin, direct, Bilirubin, total, Blood, urine, C reactive protein, Calcium, total, Chloride, Cholesterol, HDL, Cholesterol, LDL, Cholesterol, total, Creatinine, D-dimer, EGFR, Eosinophil (absolute), Eosinophil (percent), Ferritin, Glucose, HCO3, HCT, HGB, HbA1c, IG (absolute), IG (percent), INR, LE (urine), Lactate, Lipase, Lymphocyte (absolute), Lymphocyte (percent), MCH, MCHC, MCV, MPV, Magnesium, Monocyte (absolute), Monocyte (percent), Neutrophil (absolute), Neutrophil (percent), PLT, PTT, Phosphorous, Potassium, Protein, plasma, Protein, urine, RBC, RCW_CV, RCW_SD, Sodium, Specific gravity, urine, Triglycerides, Troponin T, Urea nitrogen, Urobilinogen, urine, WBC, aPTT, nRBC (absolute), pH, urine
Related	Physiological measurements explicitly relating to respiratory support	Baseline	**(median)** FiO2, O2 flow rate
O2 Del Method	EHR-native representation of respiratory support	Baseline, Related	**(presence and last)[b]** CPAP, ETT, Face tent, High flow nasal cannula, High humidity nasal cannula i.e.,.Optiflow, NPPV, Nasal cannula, Non-rebreather mask, Other, Simple mask, T-Piece, Trach mask/collar, Trach tube, Venturi mask
Raw	Based on proposed heuristics but not assembled into episodes	Baseline, Related	**(presence and last)[b]** ECMO, HFOT, IMV, LFOT, NIMV
Proposed	Based on proposed heuristics and assembled into episodes	Baseline, Related	**(Respiratory support duration [hours] and last)[b]** ECMO, HFOT, IMV, LFOT, NIMV

[a] Feature sets, in addition to the list of features in the "features" column, also include features for the feature sets in the "included" column.

[b] Presence indicates the presence of variable within the observation window whereas last indicates the last observed respiratory support modality within the observation window.

Abbreviations: BMI, body mass index; MAP, mean arterial pressure; GCS, Glasgow Coma Scale; ALP, alkaline phosphatase; ALT, alanine aminotransferase; ASP, aspartate aminotransferase; HDL, high-density lipoprotein; LDL, low-density lipoprotein; EGFR, estimated glomerular filtration rate; HCO3, bicarbonate; HCT, hematocrit; HGB, hemoglobin; LE, leukocyte esterase; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; MPV, mean platelet volume; PLT, platelet count; PTT, prothrombin time; RBC, red blood cell count; RCW, red blood cell distribution width; CV, coefficient of variation; SD, standard deviation; WBC, white blood cell count; aPTT, activated partial thromboplastin time; nRBC, nucleated red blood cell count; FiO2, fraction of inspired oxygen; CPAP, continuous positive airway pressure; ETT, endotracheal tube intubation; NPPV, non-invasive positive pressure ventilation; ECMO, extracorporeal membrane oxygenation; HFOT, high flow oxygen therapy; IMV, invasive mechanical ventilation; LFOT, low flow oxygen therapy; NIMV, non-invasive mechanical ventilation.

Appendix 9. Visualization of Respiratory Support Representation Methods



A visual comparison of different sets of respiratory support features for a single encounter. Each red tick on the x-axis indicates 6 hours whereas each black tick indicates 24 hours. The first subplot includes patient location information. “Floor” includes any non-ICU and non-ER location. Compared to “Baseline” features based on demographics, common labs, and vital signs, “Related” also includes common measurements related to respiratory support i.e., fraction of inspired oxygen and o2 flow rate (2nd and 3rd subplots). In addition to “Baseline” and “Related” features, “O2 Del Method” also includes the EHR-native representation of respiratory support status (4th subplot and legend). Each vertical line indicates a timestamped documentation relating to respiratory support at the time. “Raw” features were derived from the proposed classification system and heuristics prior to assembly into episodes (5th subplot). Finally, the “Proposed” features include those derived from the proposed classification system and heuristics after assembly into episodes (6th subplot). The colors of the vertical lines in the “Raw” subplot follow the same coloring schema in the “Proposed” subplot.

Appendix 10. Hyperparameter Optimization

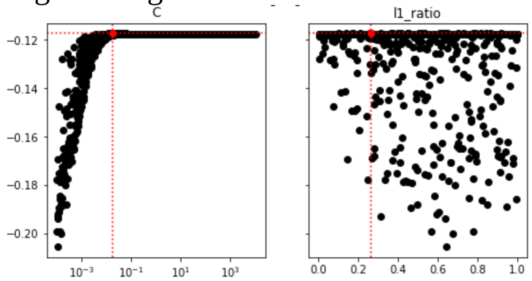
Hyperparameters for both XGB and logistic regression were optimized on the baseline set of features through 1,000 iterations of 4-fold cross-validation optimized for maximizing negative log loss. The hyperparameter optimization schema were as follows:

Model	**LogReg**	**XGB**
Preprocessing	• Normalization (zero-mean, unit-variance) • Mean imputation	None
Fixed parameters	• Penalty: Elasticnet • Solver: saga • max_iter: 1000	• tree_method: hist • single_precision_histogram: True • eval_metric: logloss
Hyperparameter search space	• C: log-uniform (1e-4, 1e4) • l1_ratio: uniform (0, 1)	• n_estimators: uniform (3, 100) • eta: log-uniform (1e-3, 1e0) • max_depth: uniform (3, 100) • min_child_weight: log-uniform (1e-2, 1e2) • gamma: log-uniform (1e-4, 1e4)

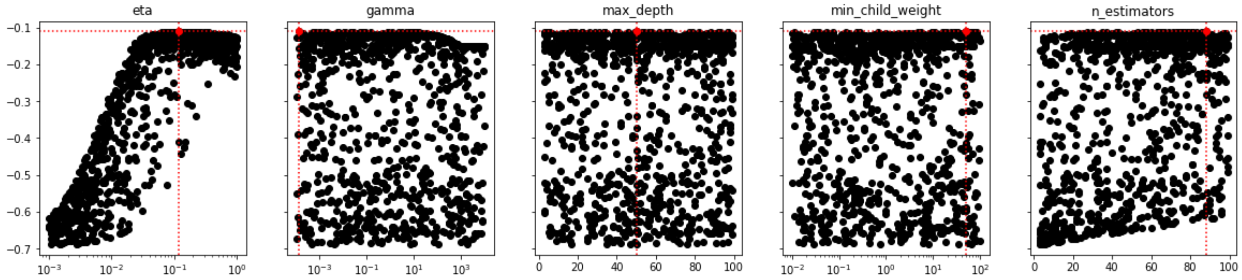
Parameter names follow the scikit-learn convention for logistic regression, and the scikit-learn API for XGB.

The relationship between parameter values and optimized metric (negative log loss) are shown as follows:

Logistic Regression:



XGB:



The optimal parameters were found to be as follows:
Logistic Regression: C: 0.0177, l1_ratio: 0.265
XGB: eta: 0.120, gamma: 0.000133, max_depth: 50, min_child_weight: 50.8, n_estimators: 88

Appendix 11. Feature Set Performances

Model	Feature Set[a]	Performance Metric		
		AUROC	AUPRC	Negative Log Loss
XGB	Baseline	0.876 (0.875 - 0.879)	0.284 (0.282 - 0.288)	-0.112 (-0.112 - -0.111)
	Related	0.885 (0.882 - 0.886)	0.317 (0.313 - 0.322)	-0.109 (-0.109 - -0.108)
	O2 Del Method	0.887 (0.884 - 0.890)	0.323 (0.317 - 0.326)	-0.108 (-0.108 - -0.107)
	Raw	0.888 (0.886 - 0.891)	0.322 (0.321 - 0.329)	-0.108 (-0.108 - -0.107)
	Proposed	0.887 (0.885 - 0.891)	0.324 (0.323 - 0.329)	-0.108 (-0.108 - -0.107)
LogReg	Baseline	0.855 (0.852 - 0.855)	0.262 (0.245 - 0.268)	-0.117 (-0.118 - -0.117)
	Related	0.859 (0.858 - 0.864)	0.286 (0.274 - 0.295)	-0.115 (-0.116 - -0.115)
	O2 Del Method	0.877 (0.874 - 0.882)	0.320 (0.314 - 0.324)	-0.110 (-0.111 - -0.110)
	Raw	0.880 (0.876 - 0.883)	0.321 (0.317 - 0.328)	-0.110 (-0.110 - -0.110)
	Proposed	0.881 (0.876 - 0.884)	0.319 (0.313 - 0.325)	-0.110 (-0.110 - -0.109)

Abbreviations: LogReg, logistic regression; XGB, eXtreme Gradient Boosted trees; AUROC, area under receiver operating characteristic curve; AUPRC, area under precision recall curve.

[a] Baseline include features based on demographics, common labs, and vital signs; Related includes common measurements related to respiratory support i.e., fraction of inspired oxygen and o2 flow rate; O2 Del Method includes the EHR-native representation of respiratory support status; Raw includes features derived from the proposed classification system and heuristics prior to assembly into episodes; Proposed includes features based derived from the proposed classification system and heuristics after assembly into episodes.

Appendix 12. Pairwise Feature Set Performance Comparisons

Performance Metric	Feature Set[a]	LogReg				XGB			
		Baseline	Related	O2 Del Method	Raw	Baseline	Related	O2 Del Method	Raw
AUROC	Related	p < 0.01				p < 0.01			
	O2 Del Method	p < 0.01	p < 0.01			p < 0.01	p < 0.01		
	Raw	p < 0.01	p < 0.01	p < 0.01		p < 0.01	p < 0.01	0.080	
	Proposed	p < 0.01	p < 0.01	p < 0.01	0.230	p < 0.01	p < 0.01	0.050	1.000
AUPRC	Related	p < 0.01				p < 0.01			
	O2 Del Method	p < 0.01	p < 0.01			p < 0.01	p < 0.01		
	Raw	p < 0.01	p < 0.01	0.030		p < 0.01	p < 0.01	p < 0.01	
	Proposed	p < 0.01	p < 0.01	0.850	0.020	p < 0.01	p < 0.01	0.080	0.560
Negative Log Loss	Related	p < 0.01				p < 0.01			
	O2 Del Method	p < 0.01	p < 0.01			p < 0.01	p < 0.01		
	Raw	p < 0.01	p < 0.01	0.010		p < 0.01	p < 0.01	0.050	
	Proposed	p < 0.01	p < 0.01	p < 0.01	0.850	p < 0.01	p < 0.01	0.050	0.770

Abbreviations: LogReg, logistic regression; XGB, eXtreme Gradient Boosted trees; AUROC, area under receiver operating characteristic curve; AUPRC, area under precision recall curve.

p-values generated through 5 replicates of 2-fold cross validation and the Wilcoxon signed-rank test (paired, two-sided)

[a] Baseline include features based on demographics, common labs, and vital signs; Related includes common measurements related to respiratory support i.e., fraction of inspired oxygen and o2 flow rate; O2 Del Method includes the EHR-native representation of respiratory support status; Raw includes features derived from the proposed classification system and heuristics prior to assembly into episodes; Proposed includes features based derived from the proposed classification system and heuristics after assembly into episodes.

Appendix 13. Data Source

All data for analysis was from Barnes-Jewish Hospital (BJH), one of the fifteen hospitals owned by BJC Healthcare, a non-profit health care organization based in St. Louis, MO and affiliated with Washington University in St. Louis, St. Louis, MO. During the time period from which the data was extracted, BJH primarily used the COMPASS EHR (Allscripts Sunrise, Chicago, IL). Clinical data was first loaded into to a hospital-managed data warehouse called Health Data Core (HDC), which is primarily used for quality improvement, then was loaded into to a university-managed research data warehouse called Research Data Core (RDC). All relevant data for inpatients between 1/12012 and 6/1/2018 was extracted from RDC.

Appendix 14. Missing Data Characterization

	Scores[b]						% missing per encounter per cohort					
Variable[a]	SIRS	qSOFA	SOFA	AOD	OD	APACHE-II	Total	Sepsis-1	CMS SEP-1	Sepsis-3	CDC ASE	ICD code
Respiratory rate	■	■				■	0.04%	0.00%	0.00%	0.00%	0.00%	0.01%
Heart rate	■					■	0.03%	0.00%	0.00%	0.00%	0.00%	0.01%
Temperature	■					■	0.07%	0.06%	0.11%	0.09%	0.16%	0.12%
$PaCO_2$	■						79.61%	48.99%	35.25%	32.93%	31.34%	45.70%
WBC	■					■	2.31%	0.03%	0.09%	0.06%	0.02%	0.09%
SBP		■			■		0.05%	0.03%	0.04%	0.03%	0.05%	0.04%
PLT			■	■	■		2.32%	0.03%	0.09%	0.06%	0.02%	0.10%
Creatinine			■	■	■		2.30%	0.04%	0.07%	0.05%	0.02%	0.07%
Bilirubin			■	■			35.58%	8.45%	3.86%	6.24%	2.51%	7.27%
MAP			■		■	■	0.05%	0.03%	0.04%	0.03%	0.05%	0.04%
PF ratio			■				50.58%	24.61%	17.12%	11.77%	16.03%	24.77%
Lactate				■	■		75.98%	40.54%	26.54%	30.83%	26.37%	23.74%
PTT					■		32.33%	13.12%	6.40%	8.19%	5.39%	12.48%
INR					■		26.08%	8.83%	3.63%	5.02%	2.91%	8.61%
pH						■	79.61%	48.99%	35.25%	32.93%	31.34%	45.70%
FiO2						■	53.07%	27.73%	21.09%	15.19%	20.07%	28.12%
SpO_2							0.04%	0.01%	0.02%	0.02%	0.03%	0.04%
HCT						■	2.19%	0.03%	0.07%	0.06%	0.02%	0.09%
PaO_2						■	79.61%	48.99%	35.25%	32.93%	31.34%	45.70%
Potassium						■	2.43%	0.06%	0.09%	0.07%	0.02%	0.13%
Sodium						■	2.19%	0.04%	0.06%	0.05%	0.02%	0.06%
A-a Gradient						■	93.30%	75.00%	63.60%	63.75%	60.13%	72.78%
BUN							2.30%	0.04%	0.07%	0.05%	0.02%	0.07%
PT							26.08%	8.83%	3.63%	5.02%	2.91%	8.61%

Abbreviations: SIRS, Systemic Inflammatory Response Syndrome; SOFA, Sequential Organ Failure Assessment; qSOFA, quick SOFA; AOD, acute organ dysfunction criteria for CDC ASE; OD, organ dysfunction criteria for CMS SEP-1; ICD, International Classification of Diseases; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event; ICU, intensive care unit; MV, mechanical ventilation; $PaCO_2$, partial pressure of arterial carbon dioxide; WBC, white blood cell count; SBP, systolic blood pressure; PLT, platelet; MAP, mean arterial pressure; PF ratio, PaO2:FiO2 ratio; PTT, partial thromboplastin time; INR, international normalized ratio; FiO_2, fraction of inspired oxygen; SpO_2, oxygen saturation; PaO_2, partial pressure of oxygen; A-a gradient, alveolar-arterial gradient; BUN, blood urea nitrogen; PT, prothrombin time.

[a] Only eligible encounters were included in the missingness analysis.

[b] Filled in cells indicate that the variable in the corresponding row is used in the score in the corresponding column.

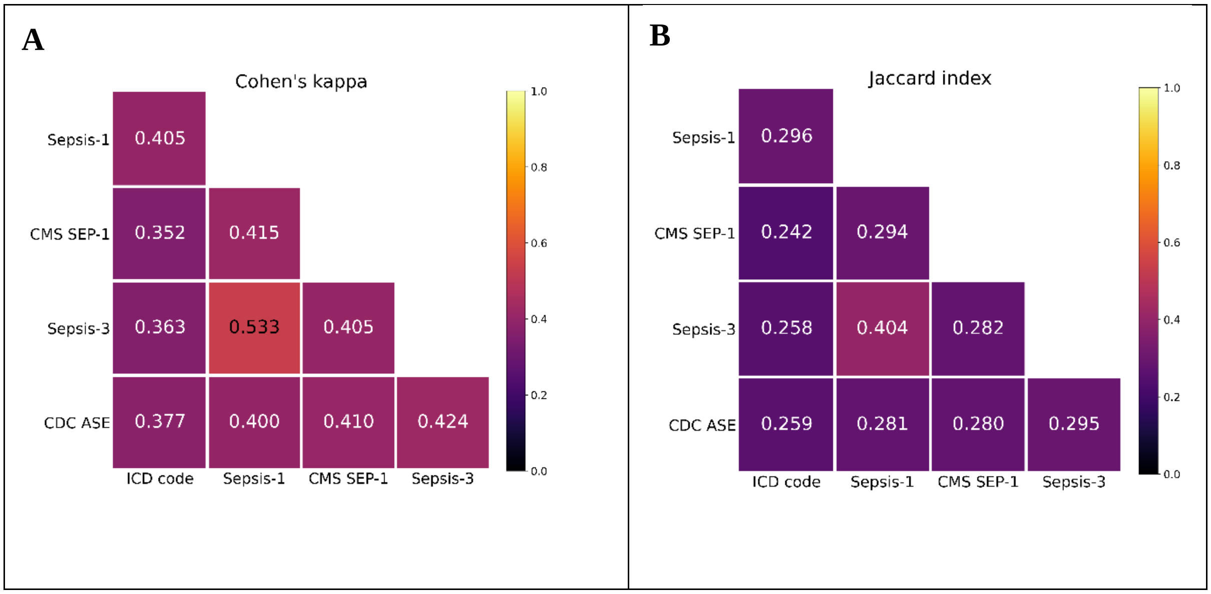
Appendix 15. Comorbidity comparison

Variable[a,b]	**Total** (n = 286.759)	**ICD code** (n = 20,670)	**Sepsis-1** (n = 32,369)	**CMS SEP-1** (n = 13,869)	**Sepsis-3** (n = 21,550)	**CDC ASE** (n = 12,494)
AIDS/HIV	2,339 (0.8%)	276 (1.3%)	458 (1.4%)	149 (1.1%)	248 (1.2%)	175 (1.4%)
Chronic pulmonary disease	69,890 (24.4%)	5,710 (27.6%)	10,109 (31.2%)	4,187 (30.2%)	7,719 (35.8%)	3,856 (30.9%)
Congestive heart failure	57,549 (20.1%)	5,766 (27.9%)	8,496 (26.2%)	4,443 (32.0%)	7,839 (36.4%)	4,014 (32.1%)
Diabetes	77,623 (27.1%)	6,603 (31.9%)	9,696 (30.0%)	4,270 (30.8%)	7,111 (33.0%)	4,011 (32.1%)
Hypertension	116,953 (40.8%)	8,034 (38.9%)	12,973 (40.1%)	5,567 (40.1%)	9,365 (43.5%)	5,179 (41.5%)
Hypothyroidism	34,417 (12.0%)	2,860 (13.8%)	4,256 (13.1%)	1,940 (14.0%)	3,365 (15.6%)	1,863 (14.9%)
Liver disease	21,942 (7.7%)	3,283 (15.9%)	4,148 (12.8%)	2,595 (18.7%)	3,464 (16.1%)	2,198 (17.6%)
Peripheral vascular disorders	21,372 (7.5%)	2,008 (9.7%)	2,653 (8.2%)	1,329 (9.6%)	2,343 (10.9%)	1,204 (9.6%)
Pulmonary circulation disorders	9,473 (3.3%)	1,239 (6.0%)	1,899 (5.9%)	981 (7.1%)	1,649 (7.7%)	746 (6.0%)
Chronic renal failure	57,021 (19.9%)	5,816 (28.1%)	8,074 (24.9%)	3,934 (28.4%)	6,519 (30.3%)	3,578 (28.6%)
Cancer	50,035 (17.4%)	4,193 (20.3%)	6,940 (21.4%)	2,838 (20.5%)	4,006 (18.6%)	2,447 (19.6%)

[a] Derived from Elixhauser comorbidities.[107, 108]

[b] Comparison across all four definition-based cohorts was performed using the χ^2 test. All comparisons were statistically significant (p<0.01).

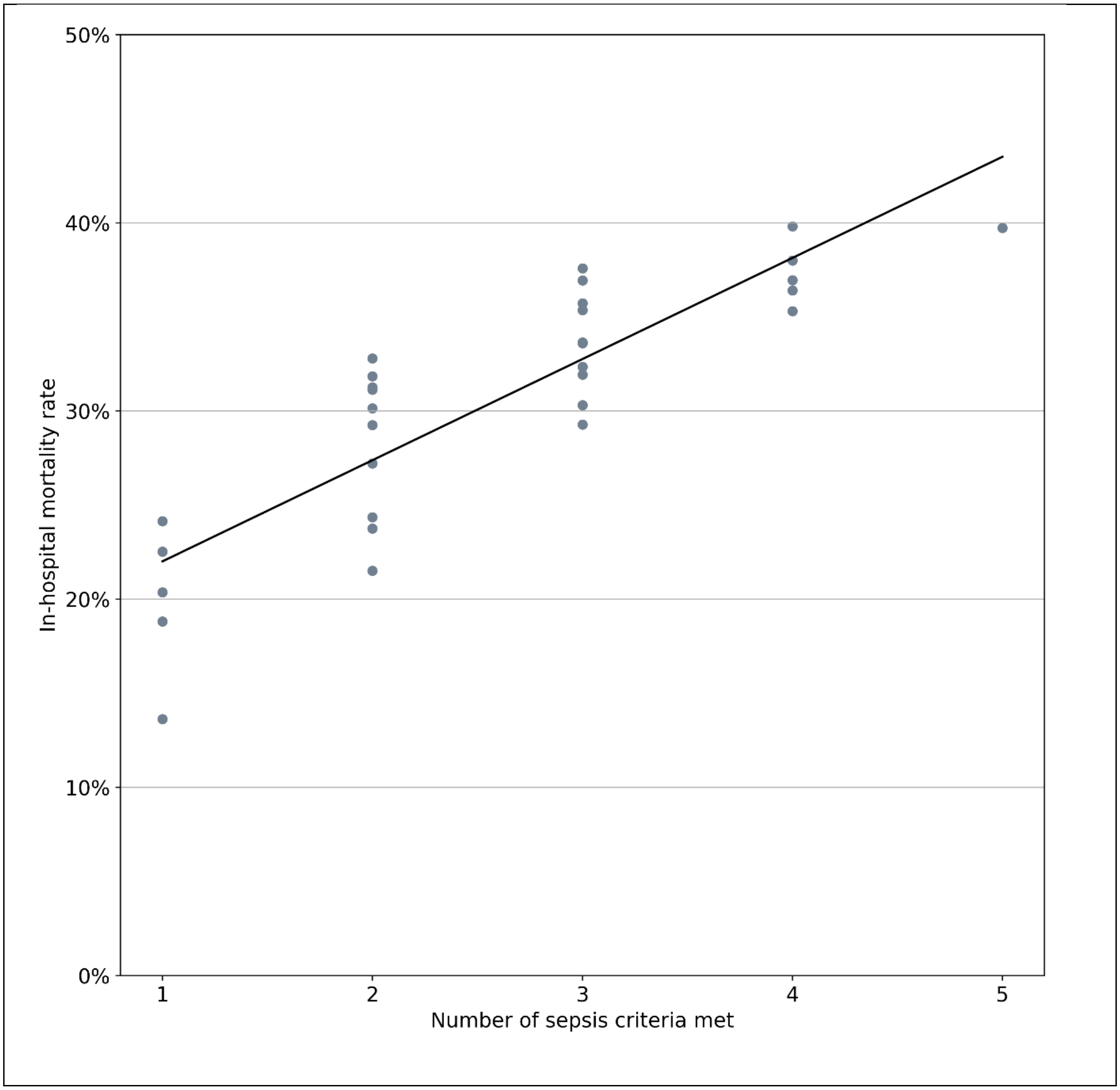
Appendix 16. Pairwise Agreement Among Sepsis Definitions



A, Cohen’s kappa and **B**, Jaccard index for each sepsis definition pair. Cohen’s kappa is a measurement of inter-rater reliability that takes into account the probability of chance agreement. Jaccard index is the size of the intersection divided by the size of the union.

Abbreviations: ICD, International Classification of Diseases; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event.

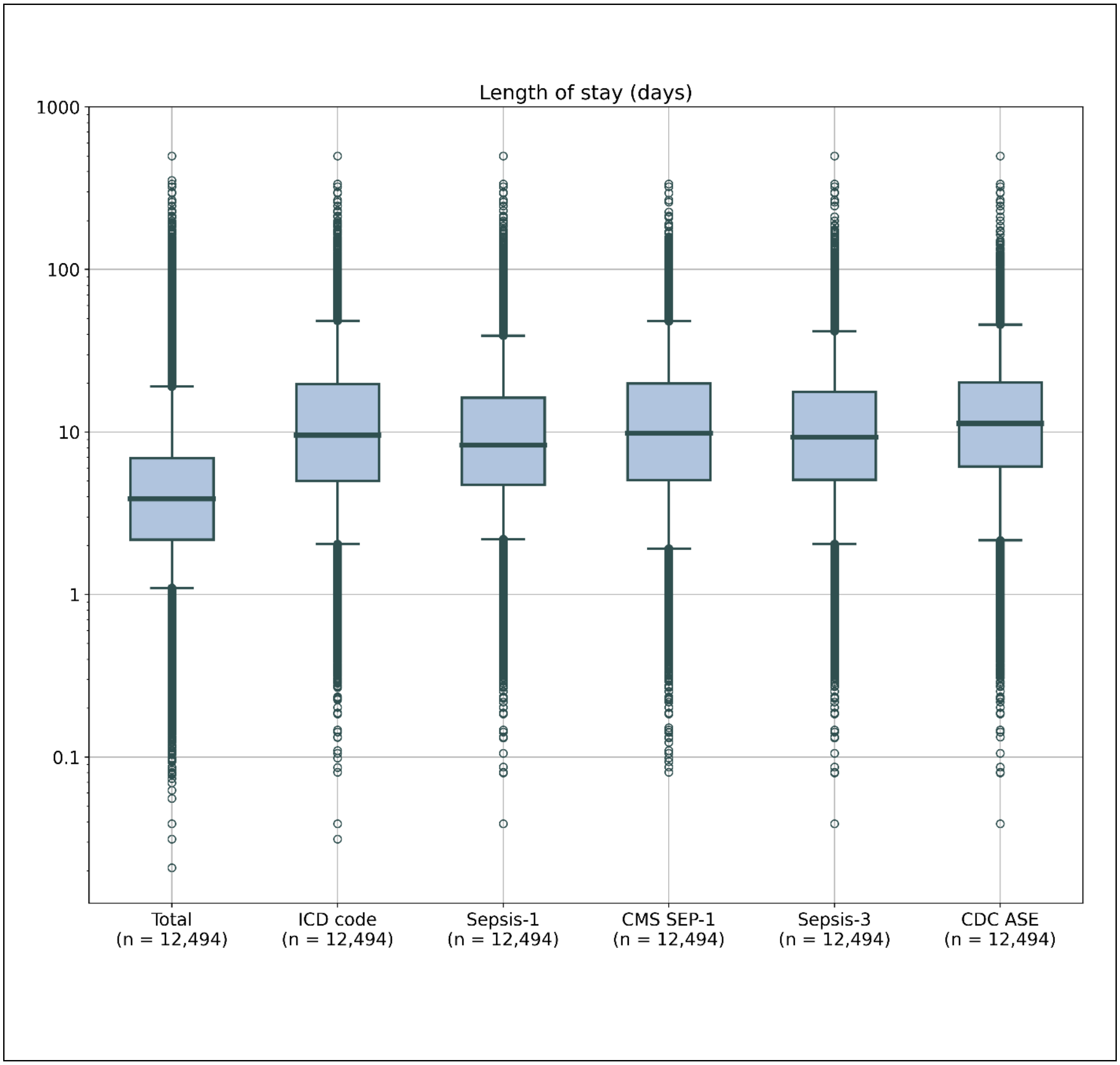
Appendix 17. Mortality by Fulfilled Number of Sepsis Definitions



Considered sepsis criteria include: ICD billing code, Sepsis-1, CMS SEP-1, Sepsis-3, and CDC ASE criteria. Patients meeting all five criteria had an in-hospital mortality rate of 39.7%. Least squares regression equation is: [In-hospital mortality rate] = 0.167 + 0.054 * [Number of sepsis criteria met]. The regression r^2 value is 0.740, and the coefficient for the number of sepsis criteria met is significant (p<0.01).

Abbreviations: ICD, International Classification of Diseases; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event; APACHE II, Acute Physiology and Chronic Health Evaluation; SOFA, Sequential Organ Failure Assessment.

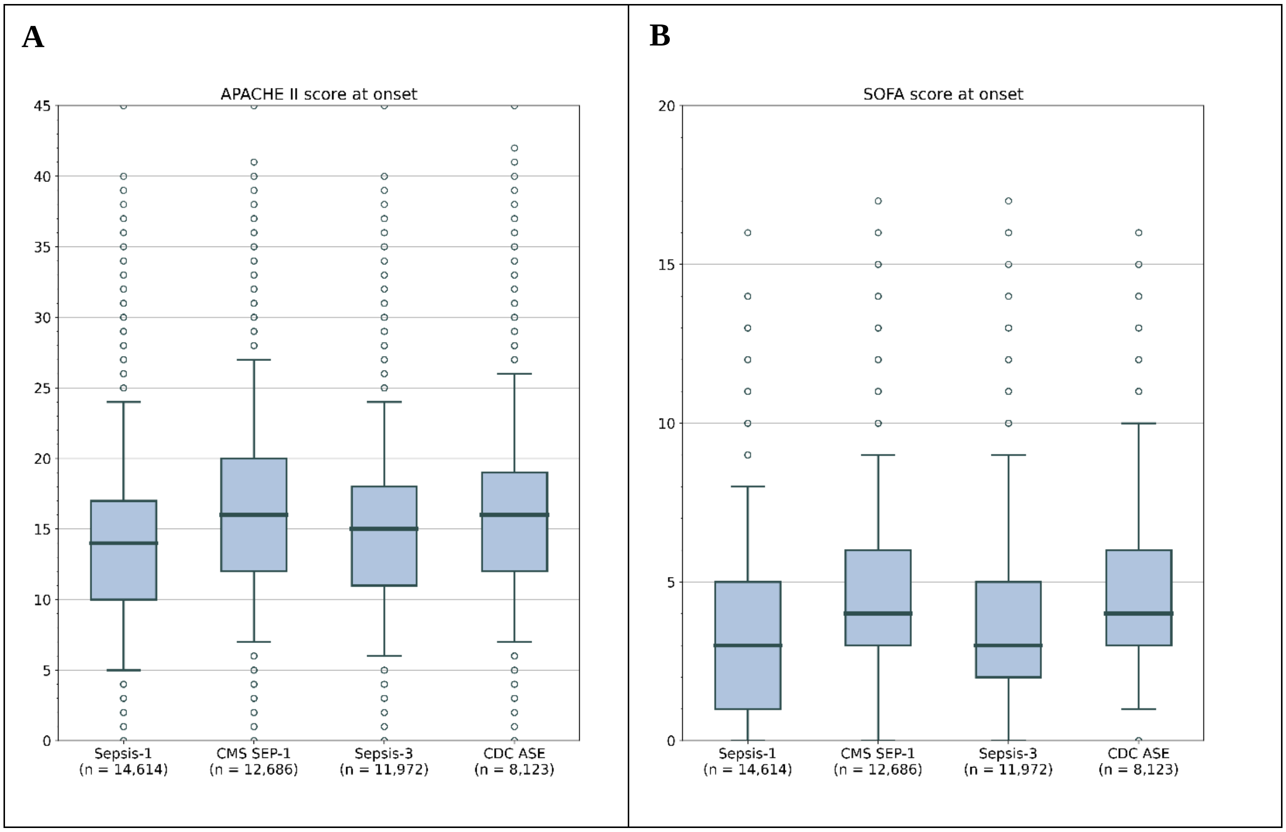
Appendix 18. Length of Stay Comparison



Hospital length of stay stratified by sepsis definition. *p*<0.01 between all definitions. Whiskers represent the 5th and 95th percentile.

Abbreviations: ICD, International Classification of Diseases; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event.

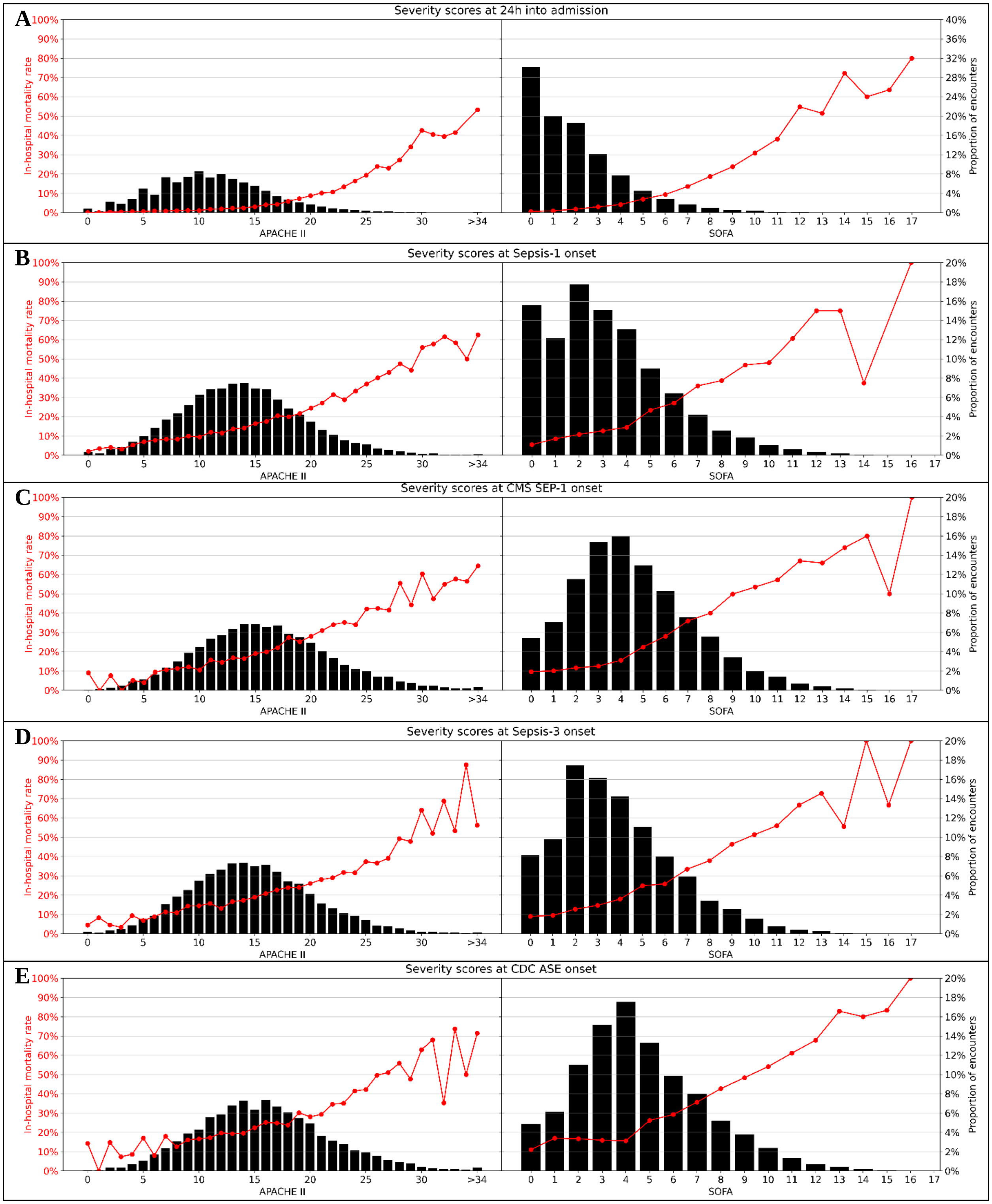
Appendix 19. Illness Severity at Onset Comparison



A, APACHE-II and **B**, SOFA score at time of sepsis onset, stratified by definition. p<0.01 between all definitions for both A and B. Scores were only calculated for those with sufficient data, which was defined as at least one measurement of each of the following within the 24h preceding sepsis onset: temperature, heart rate, respiratory rate, blood pressure, SpO_2, creatinine, and white blood cell count. Whiskers represent the 5th and 95th percentile.

Abbreviations: CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event; APACHE, Acute Physiology And Chronic Health Evaluation; SOFA, Sequential Organ Failure Assessment.

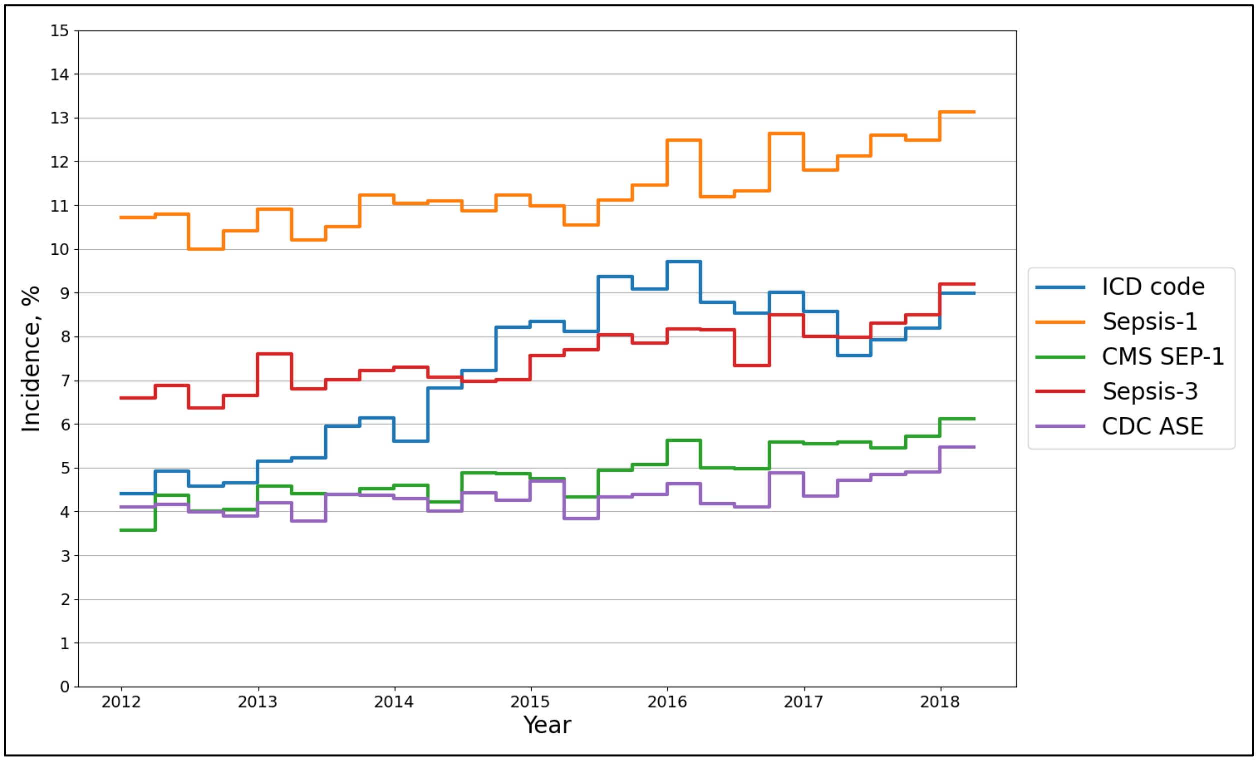
Appendix 20. Association between Illness Severity Score and In-hospital Mortality.



A, illness severity score for all eligible encounters with a length of stay of at least 24 hours (n = 276,467). **B-E,** illness severity score for sepsis cohorts at time of onset. Illness severity scores were calculated for patients in both the ICU and non-ICU settings.

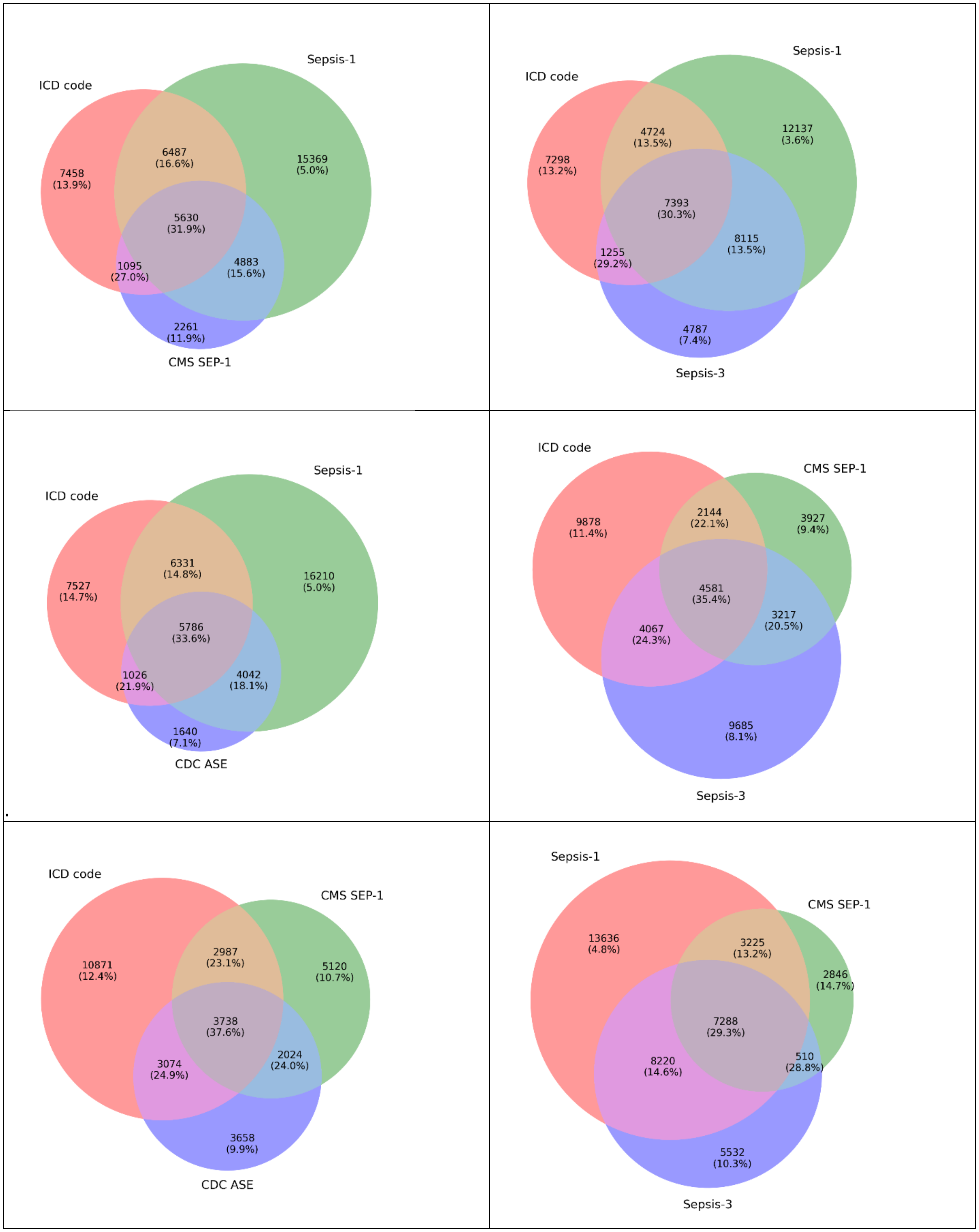
Abbreviations: CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event; APACHE II, Acute Physiology and Chronic Health Evaluation; SOFA, Sequential Organ Failure Assessment.

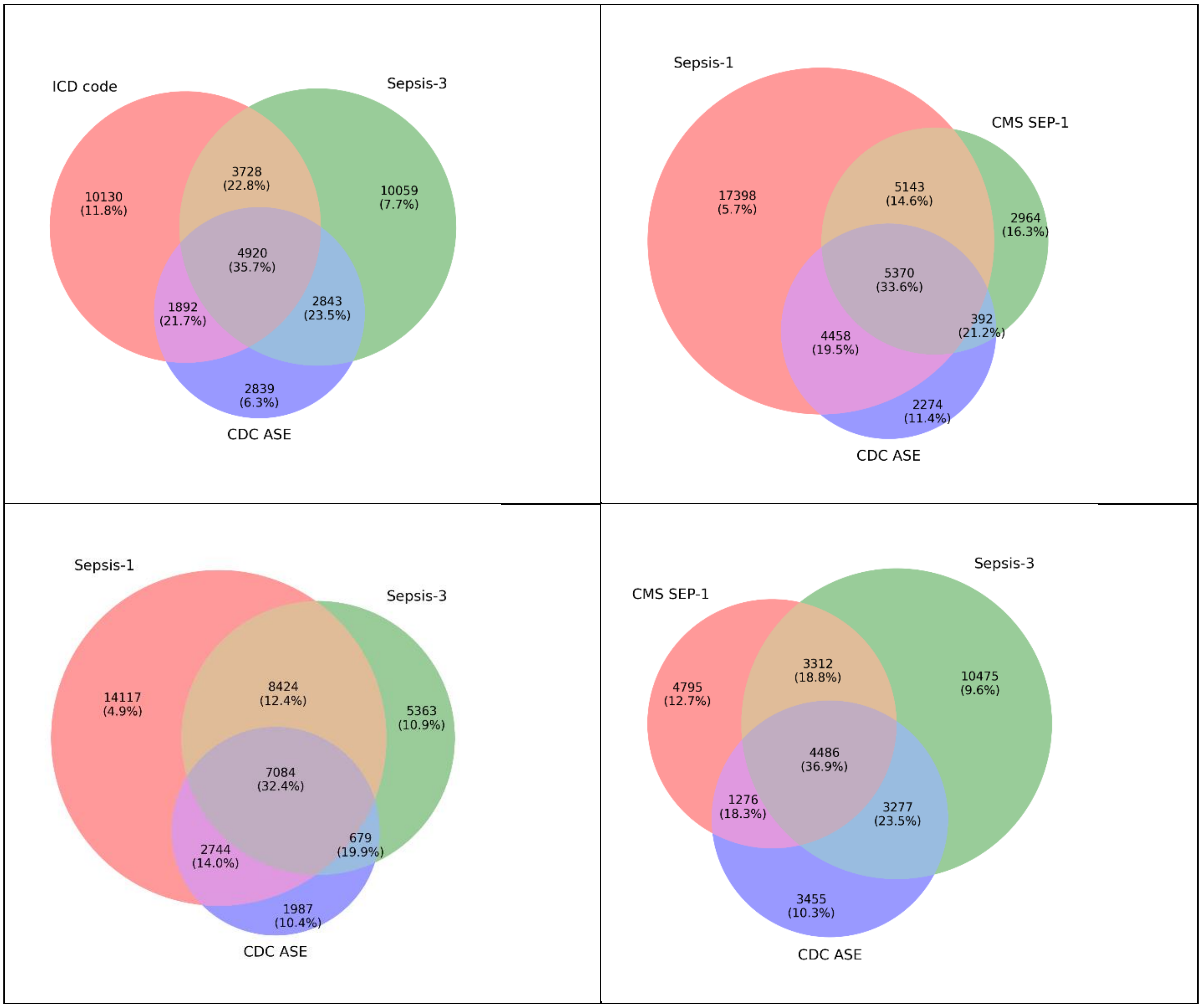
Appendix 21. Sepsis Incidence Over Time



Abbreviations: CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event.

Appendix 22. Definition Triplet Venn Diagrams





Venn diagram for each of the 10 possible triplet combinations of compared sepsis definitions. Percentages in parenthesis indicate in-hospital mortality rate.

Abbreviations: ICD, International Classification of Diseases; CMS SEP-1, Centers for Medicare and Medicaid Services severe sepsis core measure 1; CDC ASE, Centers for Disease Control and Prevention Adult Sepsis Event.

Appendix 23. Overview

In order to codify and implement the established definitions for sepsis, the clinical descriptions must be converted into computable forms, which includes an inherent degree of subjectivity. In this section, the objective is to elucidate the precise means through which the various definitions were implemented – either faithfully or with prudent modification. The most elusive and controversial is the concept of suspected or confirmed infection. Infection status was meant to be assessed at the bedside (as in Sepsis-1) or determined via manual chart review (as in CMS SEP-1). Seymour defines suspicion of infection as concomitant cultures and antibiotics which has the benefit of being practical and explicit, but has invited criticism for being tautological due to the use of physician actions in the diagnostic criteria.[120] Regardless, because the Seymour suspicion of infection criteria can be executed easily at scale using EHR data, we adopted the structure of the Sepsis-3 definition – a dyad of (suspected) infection and response (to infection) – for Sepsis-1 and CMS SEP-1. Other modulations of the criteria that can have a significant impact on the resulting cohort includes how often one can "re-litigate" a patient encounter for sepsis; in other words, if one should one investigate only the first episode of infection or every episode of infection during a hospital encounter. Seymour et al. (Sepsis-3) indicate that "only the first episode of suspected infection for each encounter" was considered whereas Rhee et al. (CDC ASE) explain that "multiple windows during a hospitalization are possible."[127, 128] All anti-infectives and cultures were mapped by both a clinical pharmacist and critical care physician. Disagreement was adjudicated by a third critical care physician with over 20 years of critical care experience.

Appendix 24. Clinical Data Preprocessing and Mapping

Raw clinical data were mapped to cogent clinical concepts through a combination of informatics approaches and subject matter expert manual review. Certain data elements were not present or partially present, but were able to be derived from related data elements:

- BMI = weight (kg) / (height (m))^2. BMI was explicitly present for 35.3% of the study population, was able to be calculated for 91.8%, and was ultimately available for 92.0%.
- FiO2 was available explicitly, but was also calculated whenever there was oxygen flow documentation according to the following formula: oxygen flow x 3.5 + 21.
- PaO2 - FiO2 ratio (PFRatio) was calculated whenever there was documentation of either PaO2 or FiO2. From each documentation, we looked back 24 hours for the latest complement documentation (PaO2 for FiO2 and vice versa) to calculate the ratio. If a complement FiO2 could not be found for PaO2, FiO2 was assumed to be 21%. If a complement PaO2 could not be found for PaO2, PaO2 was calculated using the following formula: 100 – Age (years) * 0.3
- Estimated glomerular flow rate (eGFR) was calculated according to the MDRD study equation: 175 * Creatinine$^{-1.154}$ * Age$^{-0.203}$ * ((Gender == Female)*.742)) * ((Race==Black)*1.212)
- Blood urea nitrogen – creatinine ratio (BUNCr ratio) was calculated whenever there was a blood urea nitrogen documentation and creatinine documentation within a one-hour window as blood urea nitrogen / creatinine. Time of documentation was set as the later of the two.

- Shock index (SI) was calculated whenever there was a heart rate documentation and a systolic blood pressure documentation within a one-hour window as heart rate / systolic blood pressure. Time of documentation was set as the later of the two.

Appendix 25. Detailed criteria for ICD-code surveillance definition.

Condition	ICD version	ICD code list[a]	Admitting diagnosis	Discharge diagnosis
Sepsis	9	995.91, 038.9, 038.0, 038.10, 038.11, 038.12, 038.19, 038.2, 038.40, 038.41, 038.42, 038.43, 038.44, 038.49, 038.8	3,620	11,548
	10	A41.9, A40.9, A41.2, A41.01, A41.02, A41.1, A40.3, A41.4, A41.50, A41.3, A41.51, A41.52, A41.53, A41.59, A41.89, A02.1, A22.7, A26.7, A32.7, A40.0, A40.1, A40.8, A41.81, A42.7, A54.86, B37.7	3,869	14,278
Severe Sepsis	9	995.92	10	6,702
	10	R65.20	49	2,105
Septic Shock	9	785.52	15	4,379
	10	R65.21	109	5,414

Abbreviations: ICD, International Classification of Diseases.

[a] Sepsis ICD diagnosis code list from Buchman, et al.[117]

Appendix 26. Detailed Criteria for Sepsis-1 Surveillance Definition

1) Overview: Sepsis-1 is defined as the systemic inflammatory response to infection.[121] The Sepsis-1 clinical criteria is comprised of systemic inflammatory response syndrome (SIRS) criteria and suspicion of infection.

2) Infection (suspected, presumed, and/or confirmed): The source guideline is unspecific on how to surveil for infection status. In order to implement the definition in a computable manner, we adopted the Sepsis-3 conceptual framework for suspicion of infection which is concomitant cultures and antibiotics.[121] Suspicion of infection is defined as cultures followed by antibiotics within 48 hours or antibiotics followed by cultures within 24 hours. Time of suspicion of infection is the earlier of the two – either antibiotic order start time or culture collection time. Consecutive antibiotic orders were merged as a single antibiotic regimen with a tolerance of 1 day, and only antibiotic regimens of at least 2 qualifying antibiotic days were considered. If the patient died or was discharged to hospice or an acute care hospital, antibiotic regimens leading up to the end date of the encounter could qualify.

2.1) Antimicrobials: All intravenous antibiotics, antivirals and antifungals were included with the addition of enteral vancomycin and metronidazole to account for the practice changing treatment guidelines for C. diff released in 2018.[194]

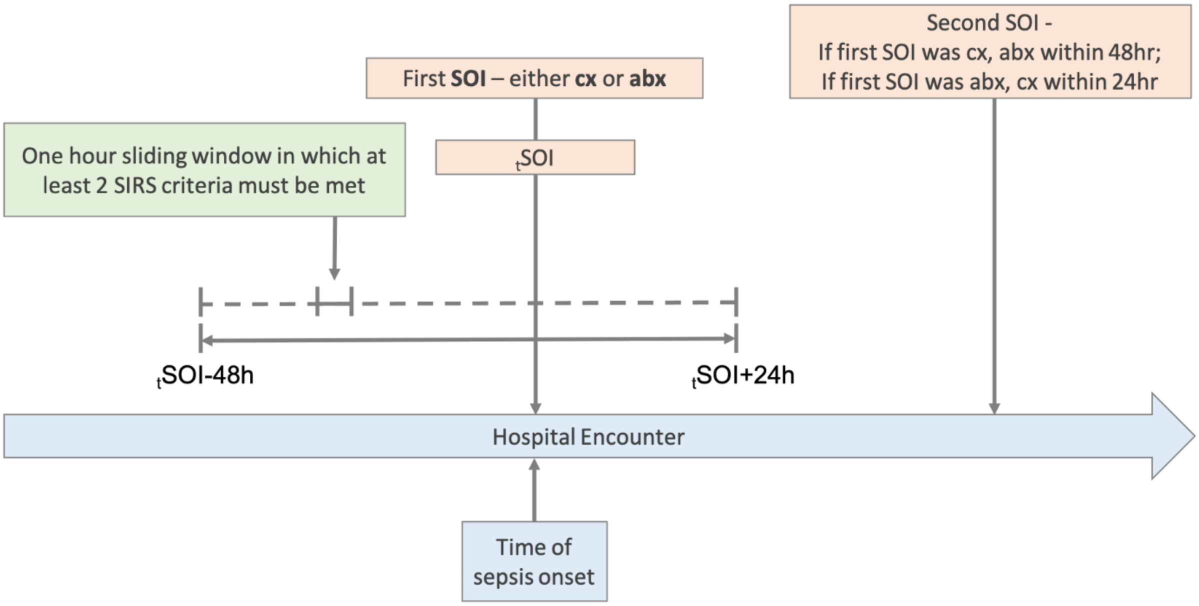
2.2) Cultures:

TESTS INCLUDED AS “CLINICAL CULTURES”[128]
Bacterial and fungal cultures from the following sites: blood, urine, respiratory tract, cerebrospinal fluid (CSF), medical devices, catheter tips, bile, pleural, peritoneal, joint/synovial, wound, drain, cyst, sinus, abscess.
Respiratory viral tests: respiratory viral multiplex, influenza swabs, adenovirus.
C. diff assays: ELISA, PCR.
Specific organism antigens from serum, urine, or CSF, such as: Histoplasma, Blastomycosis, Cryptococcus, Coccidioidomycosis, Legionella.
Specific organism cultures or smears: Pneumocystis, Legionella.
Specific organism PCRs: Ehrlichia, CMV, Toxoplasma, Borrelia, Mycoplasma.
Tests not included: Gram stain without culture, tests for parasites and acid-fast bacilli, sexually transmitted infections, serological tests (IgM, IgG), surveillance cultures (e.g., MRSA, VRE), HIV, hepatitis, H. Pylori, fungal markers, hepatitis.

3) Response to infection: SIRS criteria were met if 2 or more of the following are met: temperature >38.0 C or <36.0 C; heart rate >90; respiratory rate >20 per minute; white blood cell count >12,000 or <4,000 or >10% bands. The source guidelines are unspecific on the time window within which at least 2 SIRS criteria subcomponents must be met, so it was restricted it to one hour.

4) Sepsis: Encounters were identified as sepsis cases if SIRS criteria were met within a time window (48h before to 24h after) surrounding time of suspected infection. Time of sepsis onset was defined as time of suspected infection.

Sepsis-1


Second SOI -
If first SOI was cx, abx within 48hr;
If first SOI was abx, cx within 24hr
First **SOI** – either **cx** or **abx**
One hour sliding window in which at least 2 SIRS criteria must be met
tSOI
tSOI-48h
tSOI+24h
Hospital Encounter
Time of sepsis onset

Abbreviations: SIRS = systemic inflammatory response syndrome; SOI = suspicion of infection; tSOI = time of suspicion of infection; abx = antibiotics; cx = cultures.

Appendix 27. Detailed Criteria for CMS SEP-1 Surveillance Definition

1) Overview: In 2015, the Centers for Medicare and Medicaid Services (CMS) established a series of quality metrics that standardize the criteria for severe sepsis and septic shock recognition.[126] These criteria rely on a triad of suspicion of infection, the presence of SIRS criteria, and organ dysfunction.

2) Infection (suspected, presumed, and/or confirmed): According to the specification manual, a wide variety of clinical documentation could serve as evidence of infection. However, many of the documentation types rely on free text data elements (e.g., clinical notes) and were intended for manual abstraction. Per the CMS-criteria, "If an antibiotic is ordered for a condition that may be inflammation or a sign or symptom of an infection this may be considered documentation of an infection." As a surrogate for clinical documentation, a sepsis-relevant antibiotic order was considered to be evidence for suspicion of infection. Consecutive antibiotic orders were merged as a single antibiotic regimen with a tolerance of 1 day, and only antibiotic regimens of at least 2 qualifying antibiotic days were considered. If the patient died or was discharged to hospice or an acute care hospital, antibiotic regimens leading up to the end date of the encounter could qualify.

2.1) Antimicrobials: Because the specification manual specifically states to exclude documentation of viral, fungal, and parasitic infections, all intravenous antibiotics (excepting antivirals, antifungals, or antiparasitics) were included with the addition or enteral vancomycin and metronidazole to account for the practice changing treatment guidelines for C. diff released in 2018.[194]

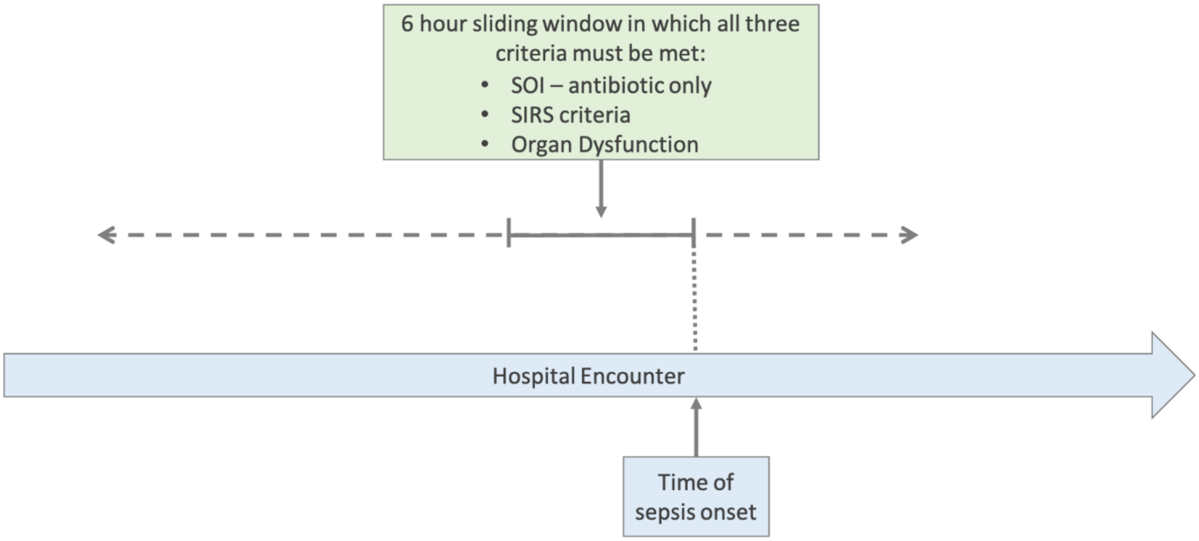
2.2) Cultures: N/A

3) Response to infection: SIRS criteria are met if 2 or more of the following are met: temperature >38.0 C or <36.0 C; heart rate >90; respiratory rate >20 per minute; white blood cell count >12,000 or <4,000 or >10% bands. Organ dysfunction criteria required at least one of the following:

- Hypotension as documented by a systolic blood pressure (SBP) <90 mmHg or mean arterial pressure <65 mmHg, excluding orthostatic BP evaluation or SBP decrease of more than 40 mmHg. Since the drop of blood pressure by more than 40 required knowledge of a patient's baseline or documentation from a treating clinician, these were omitted. Orthostatic blood pressures were also merged into all blood pressure readings since their presence in the data set was so scarce.
- Acute respiratory failure as documented by the need for new invasive or non-invasive mechanical ventilation.
- Acute kidney injury defined as either a rise in serum creatinine (Cr) by 0.5 from baseline, a Cr ≥ 2, excluding those who had end-stage renal disease (ESRD), or oliguria defined by a urine output < 0.5 mL/kg/hour for 2 consecutive hours. ESRD was identified by ICD codes: N17* or 584*. Since baseline Cr and hourly urine output was not available, these were omitted.
- Acute hepatic injury as defined by a total bilirubin >2 mg/dL.
- Platelet count <100,000 cells/µL.
- INR >1.5 or aPTT >60 sec. Outpatient medications were not a part of this dataset, so it was assumed that all patients were not on anticoagulation.
- Lactate >2 mmol/L.

4) Sepsis: In accordance with the guidelines, all criteria had to be met within 6-hours and the time of onset was defined as the time the last of the criteria were met (below).

CMS SEP-1



Abbreviations: SIRS = systemic inflammatory response syndrome; SOI = suspicion of infection; tSOI = time of suspicion of infection; abx = antibiotics; cx = cultures.

Appendix 28. Detailed Criteria for Sepsis-3 Surveillance Definition

1) Overview: In 2016, Sepsis-3, the third international consensus definition of sepsis, was released and simultaneously evaluated.[118, 127] Sepsis was then defined as organ dysfunction caused by a dysregulated host response to infection.

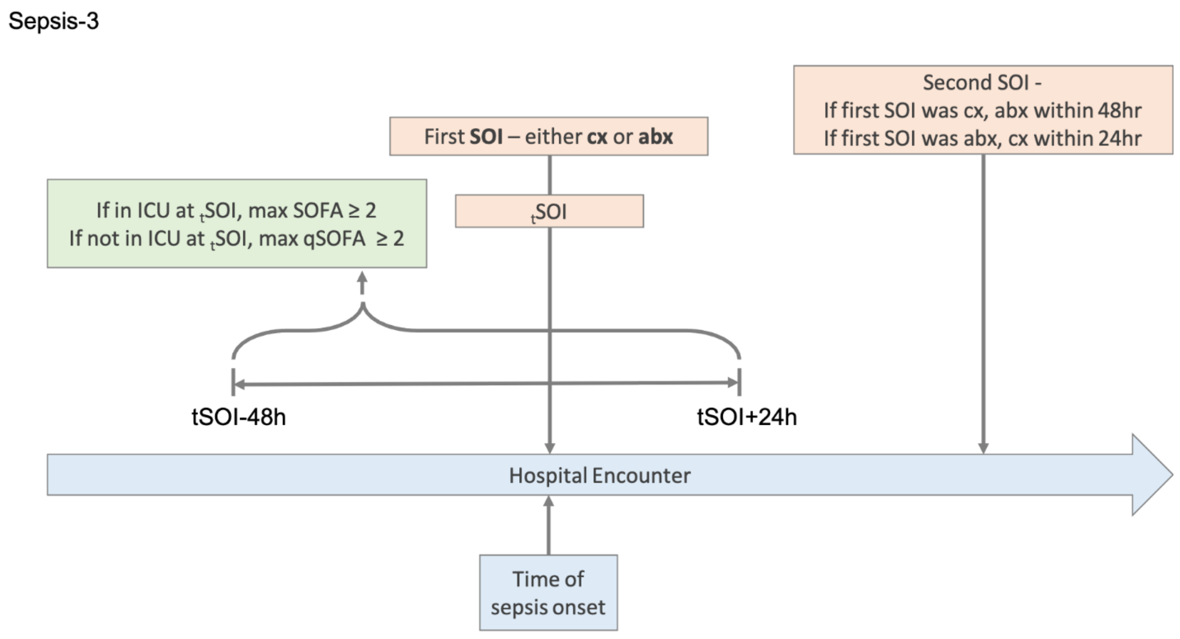
2) Infection (suspected, presumed, and/or confirmed): Infection according to the Sepsis-3 criteria required antibiotics within 72 hours of culture or a culture within 24 hours of antibiotic administration. Only the first episode of suspected infection was identified for each encounter, as described in the supplementary material (eAppendix A) in the source publication. Time of suspected infection was set as the earlier of either culture collection time or antibiotic order start time.

2.1) Antimicrobials: Consistent with the source publication, all oral and IV antibiotics were included and one time perioperative antibiotics were excluded.

2.2) Cultures: Consistent with the source publication, all bacterial, fungal, viral and parasitic cultures as well as C. diff assays from the following sites were included: abdomen, bronchoalveolar lavage, blood, bone, cerebral spinal fluid, catheters/devices, pleural space, skin/tissue, stool, urinary tract.

3) Response to infection: Authors of the source publication explain that while both the Sequential Organ Failure Assessment (SOFA) and qSOFA scores are primarily recommended for sepsis-categorization, qSOFA has the benefit of not requiring laboratory measurements and can be assessed quickly and repeatedly, which is especially useful in a non-ICU setting. Moreover, Seymour et al. noted that qSOFA has a "predictive validity outside of the ICU that is statistically greater than the SOFA score." We thus decided to use SOFA in the ICU setting and qSOFA in the non-ICU setting. Consistent with the source publication, the thresholds for both qSOFA and SOFA were set at ≥ 2.

4) Sepsis: Consistent with the source publication, encounters were identified as sepsis cases if the response to infection criteria was met within a time window (48h before to 24h after) surrounding time of suspected infection. If patient was in an ICU at time of suspected infection, SOFA was used; otherwise, qSOFA was used. Time of sepsis onset was defined as time of suspected infection.


Sepsis-3
Second SOI -
If first SOI was cx, abx within 48hr
If first SOI was abx, cx within 24hr
First SOI – either cx or abx
If in ICU at tSOI, max SOFA ≥ 2
If not in ICU at tSOI, max qSOFA ≥ 2
tSOI
tSOI-48h
tSOI+24h
Hospital Encounter
Time of sepsis onset

Abbreviations: SOI = suspicion of infection; tSOI = time of suspicion of infection; qSOFA = quick sequential organ failure assessment; SOFA = sequential organ failure assessment; abx = antibiotics; cx = cultures.

Appendix 29. Detailed Criteria for CDC ASE Surveillance Definition

1) Overview: In 2017, Rhee et al. published an epidemiological study investigating sepsis incidence and trends using an alternative interpretation of the Sepsis-3 definition which was later adopted by Centers for Disease Control and Prevention and named the Adult Sepsis Event.[128] Like Sepsis-3, CDC ASE criteria is comprised of signs of infection and organ dysfunction.

2) Infection (suspected, presumed, and/or confirmed): Presumed infection was defined as initiation of an antimicrobial regimen with a minimum of 4 qualifying antimicrobial days (QAD) starting within 2 days of blood culture. As described in the source publication, at least 1 antimicrobial in the first 4 QAD must be intravenous. If the patient died or was discharged to hospice or an acute care hospital, antibiotic regimens leading up to the end date of the encounter was included for analysis.

2.1) Antimicrobials: All antimicrobials explicitly listed in eAppendix B of the reference publication (Rhee, 2017) were used.

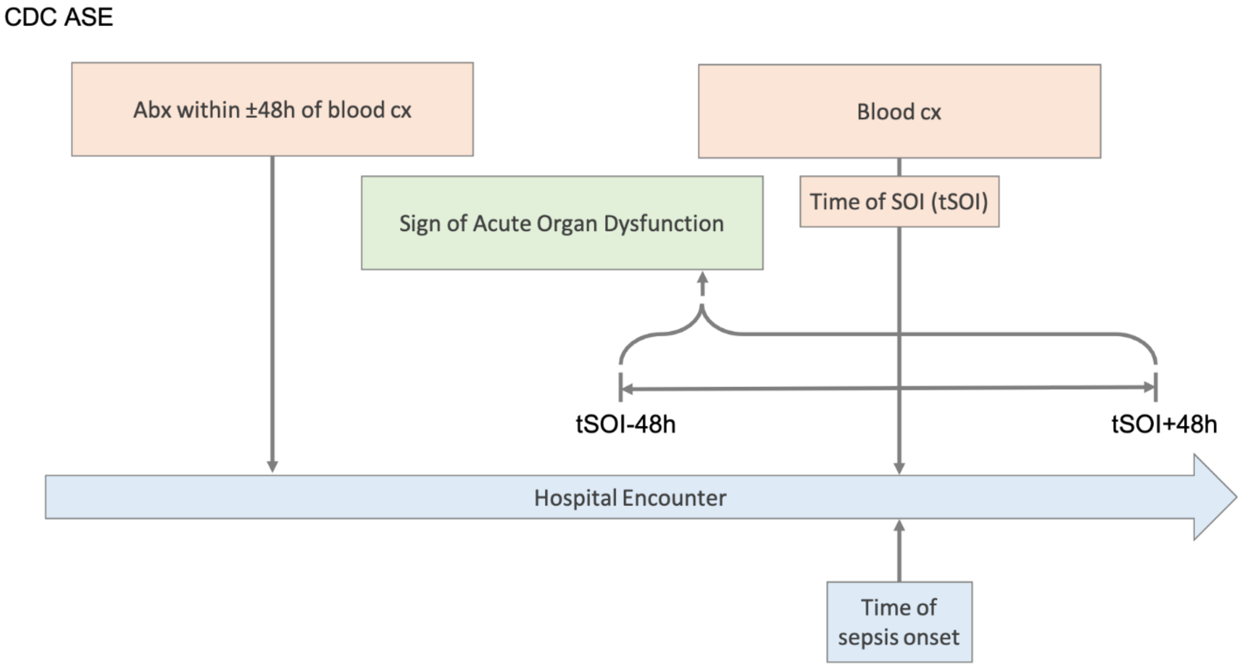
ANTIBIOTICS
IV Antibacterials
amikacin, ampicillin, ampicillin/sulbactam, azithromycin, aztreonam, cefamandole, cefazolin, cefepime, cefmetazole, cefonicid, cefoperazone, cefotaxime, cefotetan, cefoxitin, ceftaroline, ceftazidime, ceftazidime/avibactam, ceftizoxime, ceftolozane/tazobactam, ceftriaxone, cefuroxime, cephalothin, cephapirin, chloramphenicol, ciprofloxacin, clindamycin, cloxacillin, colistin, dalbavancin, daptomycin, doripenem, doxycycline, ertapenem, gatifloxacin, gentamicin, imipenem, kanamycin, levofloxacin, lincomycin, linezolid, meropenem, methicillin, metronidazole, mezlocillin, minocycline, moxifloxacin, nafcillin, oritavancin, oxacillin, penicillin, piperacillin, pileracillin/tazobactam, polymyxin B, quinupristin/dalfopristin, streptomycin, tedizolid, telavancin, ticarcillin, ticarcillin/clavulanate, tigecycline, tobramycin, trimethoprim/sulfamethoxazole, vancomycin
PO Antibacterials
amoxicillin/clavulanate, amoxicillin, ampicillin, azithromycin, cefaclor, cefadroxil, cefdinir, cefditoren, cefixime, cefpodoxime, cefprozil, ceftibuten, cefuroxime, cephalexin, cephradine, chloramphenicol, cinoxacin, ciprofloxacin, clarithromycin, clindamycin, cloxacillin, dicloxacillin, doxycycline, fidaxomicin, fosfomycin, gatifloxacin, levofloxacin, lincomycin, linezolid, metronidazole, minocycline, moxifloxacin, nitrofurantoin, norfloxacin, ofloxacin, penicillin, pivampicillin, rifampin, sulfadiazine, sulfadiazine-trimethoprim, sulfamethoxazole, sulfisoxazole, tedizolid, telithromycin, tetracycline, trimethoprim, trimethoprimsulfamethoxazole, vancomycin
IV Antifungals
amphotericin B, anidulafungin, caspofungin, fluconazole, itraconazole, micafungin, posaconazole, voriconazole
PO Antifungals
fluconazole, itraconazole, posaconazole, voriconazole
IV Antivirals
acyclovir, ganciclovir, cidofovir, foscarnet, peramivir
PO Antivirals
Oseltamivir

2.2) Cultures: In accordance to the source publication, only blood cultures were used.

3) Response to infection: Authors of the source publication modified the SOFA score to assess acute organ dysfunction. The new criteria required at least one of the following: initiation of a new vasopressor infusion (norepinephrine, dopamine, epinephrine, phenylephrine, vasopressin), at least 24 hours of mechanical ventilation, doubling of serum creatinine or drop in estimated glomerular filtration rate by at least 50% (excluding end-stage renal disease), a total bilirubin ≥

2.0 mg/dL with a doubling from baseline, platelet count <100 cells/µL with at least a 50% decline from baseline or serum lactate of at least 2.0 mmol/L (below).

4) Sepsis: In accordance to the source publication, encounters were identified as sepsis cases if they met the modified SOFA criteria during the time window (48h before to 48h after) surrounding blood culture. Multiple blood cultures could each serve as episodes of presumed infection, in accordance to the supplementary material in the originating publication (eAppendix A). Time of onset was defined as time of blood culture.



Abbreviations: SOI = suspicion of infection; tSOI = time of suspicion of infection; abx = antibiotics; cx = cultures.

Appendix 30. APACHE-II Implementation Details

APACHE-II was used to characterize severity of illness at time of sepsis onset. APACHE-II was intended to be measured 24h into ICU admission and every 24h afterwards. However, many sepsis patients had an onset-time fairly early into admission and thus had a time-to-onset well under 24h, often in non-ICU settings. Thus we decided to only calculate and compare APACHE-II scores for those who had a minimal set of measurements which we defined as: at least 1 heart rate, systolic blood pressure, temperature, respiratory rate, oxygen saturation (SpO_2), white blood cell count, and creatinine measurement in the 24 preceding onset. Depending on the cohort, roughly half the population had a minimal set of measurements, which is reflected in eFigure 3. In accordance to the source publication for APACHE-II, the most deranged measurement within the past 24h was used to calculate APACHE-II. Unlike in the source publication in which all measurements are mandatory, we assumed normal if missing.[2] Acute kidney injury (used to modulate the creatinine score) was determined based on diagnosis codes (N17* or 584*). History of immune-compromise or organ insufficiency were determined based on diagnosis codes. All patients were assumed to be non-operative or emergency post-operative patients, not elective post-operative patients.

Appendix 31. TRIPOD checklist

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Title
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Abstract
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	Introduction, paragraph 1 & 2
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	Introduction, paragraph 2
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	Methods, “Study Design, Data Sources, and Population” section, and eMethods 1
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	Methods, “Study Design, Data Sources, and Population” section
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	Methods, “Study Design, Data Sources, and Population” section
	5b	Describe eligibility criteria for participants.	Methods, “Study Design, Data Sources, and Population” section
	5c	Give details of treatments received, if relevant.	N/A
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	Methods, “Sepsis Definition” section and eMethods 3
	6b	Report any actions to blind assessment of the outcome to be predicted.	The outcome was determined in an automated fashion using consensus criteria definition, methods, “Sepsis Definition” section and eMethods 3
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	Methods, “Feature Generation and Engineering” section
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	The predictors were extracted in an automated fashion, methods, “Feature Generation and Engineering” section
Sample size	8	Explain how the study size was arrived at.	Methods, “Study Design, Data Sources, and Population” section and results, “Patient Population” section
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Methods, “Feature Generation and Engineering” section
Statistical analysis	10a	Describe how predictors were handled in the analyses.	Methods, “Feature Generation and

methods			Engineering” section
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	Methods, “Feature Generation and Engineering,” “Model Development,” and “Model Performance” sections
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	Methods, “Model Performance” and “Pseudo-Prospective Trial” sections
Risk groups	11	Provide details on how risk groups were created, if done.	N/A
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Methods, “Study Design, Data Sources, and Population” section; Results, “Patient Population” section, and eFigure 1
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	Results, “Patient Population” section, Table 1, eTable 2
Model development	14a	Specify the number of participants and outcome events in each analysis.	Results, “Patient Population” section, Table 1, eTable 2
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	Elided because too many. Subset shown in eFigure 2.
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Elided because too many
	15b	Explain how to the use the prediction model.	Discussion
Model performance	16	Report performance measures (with CIs) for the prediction model.	Bootstrapped mean and standard deviation reported in Figure 1 and eTable 4
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	Discussion, last paragraph
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	Results and Discussion
Implications	20	Discuss the potential clinical use of the model and implications for future research.	Discussion
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Supplementary digital content referred to and linked throughout manuscript
Funding	22	Give the source of funding and the role of the funders for the present study.	Title page, “Financial Support” section

Appendix 32. Data source

The data for analysis was sourced from Barnes-Jewish Hospital (BJH), one of the fifteen hospitals owned by BJC Healthcare, a non-profit health care organization based in St. Louis, MO and affiliated with Washington University in St. Louis, St. Louis, MO. During the time period from which the data was extracted, BJH primarily used the COMPASS EHR (Allscripts Sunrise, Chicago, IL). Clinical data was first loaded into to a hospital-managed data warehouse called Health Data Core (HDC), which is primarily used for quality improvement, then was loaded into to a university-managed research data warehouse called Research Data Core (RDC). All relevant data for inpatients between 1/12012 and 6/1/2019 was extracted from RDC.

Appendix 33. Data preprocessing and mapping

Raw clinical data were mapped to cogent clinical concepts through a combination of informatics approaches and subject matter expert manual review.

Certain data elements were not present or partially present, but were able to be derived from related data elements:

- BMI = weight (kg) / (height (m))^2. BMI was explicitly present for 35.3% of the study population, was able to be calculated for 91.8%, and was ultimately available for 92.0%.
- FiO2 was available explicitly, but was also calculated whenever there was oxygen flow documentation according to the following formula: oxygen flow x 3.5 + 21.
- PaO2 - FiO2 ratio (PFRatio) was calculated whenever there was documentation of either PaO2 or FiO2. From each documentation, we looked back 24 hours for the latest complement documentation (PaO2 for FiO2 and vice versa) to calculate the ratio. If a complement FiO2 could not be found for PaO2, FiO2 was assumed to be 21%. If a complement PaO2 could not be found for PaO2, PaO2 was calculated using the following formula: 100 – Age (years) * 0.3
- Estimated glomerular flow rate (eGFR) was calculated according to the MDRD study equation: 175 * Creatinine$^{-1.154}$ * Age$^{-0.203}$ * ((Gender == Female)*.742)) * ((Race==Black)*1.212)
- Blood urea nitrogen – creatinine ratio (BUNCr ratio) was calculated whenever there was a blood urea nitrogen documentation and creatinine documentation within a one-hour window as blood urea nitrogen / creatinine. Time of documentation was set as the later of the two.
- Shock index (SI) was calculated whenever there was a heart rate documentation and a systolic blood pressure documentation within a one-hour window as heart rate / systolic blood pressure. Time of documentation was set as the later of the two.

All numeric features were standardized (zero-mean and unit-variance) based on the distribution of the features in the training dataset. Time series data was summarized across various lookback time windows (3h, 6h, 12h, 24h, 48h, 96h) through the following aggregation functions: minimum, maximum, mean, skew, median, count, standard deviation, and last. No binning was performed. No boolean flag for presence/absence was generated.

Appendix 34. Data availability

	Total		Sepsis		Non-sepsis	
Labs and Vital Signs	% missing	# recorded	% missing	# recorded	% missing	# recorded
ALP	25.77	1 (0 - 2)	7.52	3 (2 - 7)	26.36	1 (0 - 2)
ALT	25.77	1 (0 - 2)	7.48	3 (2 - 7)	26.36	1 (0 - 2)
AST	26.21	1 (0 - 2)	7.75	3 (2 - 7)	26.81	1 (0 - 2)
A-a Gradient	97.71	0 (0 - 0)	78.92	0 (0 - 0)	98.32	0 (0 - 0)
Albumin	25.69	1 (0 - 2)	7.43	3 (2 - 7)	26.29	1 (0 - 2)
Anion Gap	0	4 (3 - 7)	0	14.5 (9 - 22)	0	4 (3 - 7)
BUN	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
BUN-Cr ratio	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
Base Excess	93.21	0 (0 - 0)	64.01	0 (0 - 1)	94.16	0 (0 - 0)
Basophils	5.15	3 (1 - 5)	2.27	8 (4 - 13)	5.25	3 (1 - 5)
Basophils abs	6.82	3 (1 - 5)	6.17	7 (4 - 13)	6.84	3 (1 - 4)
Bicarbonate	0	4 (3 - 7)	0	17 (10 - 25)	0	4 (3 - 7)
Bilirubin	25.71	1 (0 - 2)	7.43	3 (2 - 7)	26.3	1 (0 - 2)
Bilirubin direct	67.81	0 (0 - 1)	54.35	0 (0 - 1)	68.25	0 (0 - 1)
Calcium	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
Calcium ionized	89.88	0 (0 - 0)	56.26	0 (0 - 3)	90.97	0 (0 - 0)
Chloride	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
Cholesterol	64.71	0 (0 - 1)	57.62	0 (0 - 1)	64.94	0 (0 - 1)
Coombs	48.09	1 (0 - 1)	15.78	2 (1 - 3)	49.14	1 (0 - 1)
Creatinine	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
DBP	0	24 (14 - 50)	0	141 (79 - 222)	0	24 (14 - 46)
Eosinophils	5.04	3 (1 - 5)	2.4	8 (4 - 13)	5.13	3 (1 - 5)
Eosinophils abs	6.84	3 (1 - 5)	6.3	7 (4 - 13)	6.85	3 (1 - 5)
FiO2	61.87	0 (0 - 8)	14.91	41 (10 - 100)	63.4	0 (0 - 6)
Glucose	0.02	7 (3 - 25)	0	39 (16 - 101)	0.02	7 (3 - 23)
HCT	0	4 (3 - 7)	0	16 (9 - 24)	0	4 (3 - 7)
HDL	64.62	0 (0 - 1)	57.12	0 (0 - 1)	64.86	0 (0 - 1)
HGB	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
Heart Rate	0	28 (16 - 57)	0	162 (95 - 261.75)	0	27 (16 - 53)
HgA1C	65.89	0 (0 - 1)	67.36	0 (0 - 1)	65.84	0 (0 - 1)
INR	20.46	1 (1 - 3)	4.03	4 (2 - 8)	20.99	1 (1 - 3)
Immature Granulocyte	62.27	0 (0 - 2)	57.62	0 (0 - 5)	62.42	0 (0 - 1)
Immature Granulocyte abs	60.95	0 (0 - 2)	55.53	0 (0 - 6)	61.13	0 (0 - 2)
LDH	84.39	0 (0 - 0)	65.82	0 (0 - 1)	84.99	0 (0 - 0)

LDL	65.65	0 (0 - 1)	58.93	0 (0 - 1)	65.87	0 (0 - 1)
Lactic Acid	80.86	0 (0 - 0)	45.83	1 (0 - 3)	82	0 (0 - 0)
Lipase	84.06	0 (0 - 0)	83.41	0 (0 - 0)	84.08	0 (0 - 0)
Lymphocytes	4.29	3 (2 - 5)	1.9	9 (5 - 15)	4.37	3 (1 - 5)
Lymphocytes abs	6.48	3 (1 - 5)	6.07	8 (4 - 13)	6.49	3 (1 - 5)
MAP	0	24 (14 - 50)	0	141 (79 - 222)	0	24 (14 - 46)
MCH	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
MCHC	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
MCV	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
MPV	0.53	4 (3 - 7)	0.41	14 (9 - 22)	0.53	4 (2 - 7)
Magnesium	32.26	1 (0 - 3)	7.21	7 (3 - 14)	33.07	1 (0 - 3)
Monocytes	4.31	3 (2 - 5)	1.9	9 (5 - 14)	4.39	3 (1 - 5)
Monocytes abs	6.45	3 (1 - 5)	6.21	8 (4 - 13)	6.46	3 (1 - 5)
Neutrophils	4.29	3 (2 - 5)	1.9	9 (5 - 15)	4.37	3 (1 - 5)
Neutrophils abs	6.43	3 (1 - 5)	5.98	8 (4 - 13)	6.44	3 (1 - 5)
O2 Flow	62.2	0 (0 - 8)	15.19	38 (8 - 94)	63.73	0 (0 - 6)
PCO2	87.95	0 (0 - 0)	37.76	1 (0 - 6)	89.58	0 (0 - 0)
P-F Ratio	60.44	0 (0 - 8)	12.92	45 (11 - 107)	61.99	0 (0 - 7)
PLT	0.01	4 (3 - 7)	0	15 (9 - 22)	0.01	4 (2 - 7)
PO2	87.95	0 (0 - 0)	37.76	1 (0 - 6)	89.58	0 (0 - 0)
PT	20.46	1 (1 - 3)	4.03	4 (2 - 8)	20.99	1 (1 - 3)
PTT	27.15	1 (0 - 2)	7.43	4 (2 - 8)	27.79	1 (0 - 2)
Phosphorus	52.89	0 (0 - 2)	15.14	5 (1 - 11)	54.11	0 (0 - 2)
Plasma Protein	25.74	1 (0 - 2)	7.57	3 (2 - 7)	26.33	1 (0 - 2)
Potassium	0.17	4 (3 - 7)	0	16 (10 - 24)	0.18	4 (3 - 7)
RBC	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
RDW CV	0.02	4 (3 - 7)	0	15 (9 - 22)	0.02	4 (2 - 7)
RDW FL	57.41	0 (0 - 4)	53.67	0 (0 - 13)	57.53	0 (0 - 3)
Respiratory Rate	0	23 (13.25 - 48)	0	147 (82 - 240)	0	22 (13 - 45)
SBP	0	24 (14 - 50)	0	141 (79 - 222)	0	24 (14 - 46)
Shock Index	0	25 (15 - 53)	0	150 (86 - 242)	0	24 (14 - 49)
Sodium	0	4 (3 - 7)	0	16 (9 - 23)	0	4 (3 - 7)
SpO2	0	23 (13 - 48)	0	146 (80.25 - 235.75)	0	22 (13 - 45)
TSH	75.19	0 (0 - 0)	68	0 (0 - 1)	75.43	0 (0 - 0)
Temperature	0	19 (12 - 37)	0	88 (50 - 142)	0	18 (12 - 34)
Triglycerides	64.73	0 (0 - 1)	57.34	0 (0 - 1)	64.97	0 (0 - 1)
Troponin I	53.82	0 (0 - 2)	38.94	1 (0 - 3)	54.31	0 (0 - 2)

Urinalysis Ketones	94.18	0 (0 - 0)	91.61	0 (0 - 0)	94.26	0 (0 - 0)
Urinalysis Leukocyte Esterase	82.97	0 (0 - 0)	57.21	0 (0 - 1)	83.81	0 (0 - 0)
Urinalysis Nitrite	47.98	1 (0 - 1)	18.59	1 (1 - 2)	48.94	1 (0 - 1)
Urinalysis SpecificGravity	47.98	1 (0 - 1)	18.59	1 (1 - 2)	48.94	1 (0 - 1)
Urinalysis pH	47.99	1 (0 - 1)	18.59	1 (1 - 2)	48.94	1 (0 - 1)
WBC	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
eGFR	0	4 (3 - 7)	0	15 (9 - 22)	0	4 (3 - 7)
pH	87.95	0 (0 - 0)	37.76	1 (0 - 6)	89.58	0 (0 - 0)

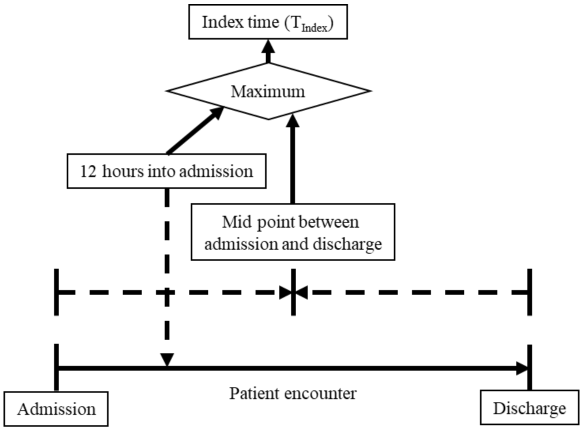
Percentage of encounters missing labs and vital signs measurements, and number of labs and vital signs measurements per encounter (median and IQR) stratified by sepsis.

Abbreviations: ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; Cr, creatinine; DBP, diastolic blood pressure; HCT, hematocrit; HDL, high density lipoprotein; HGB, hemoglobin: INR, international normalized ratio; LDH, lactate dehydrogenase; LDL, low density lipoprotein; MAP, mean arterial pressure; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; MPV, mean platelet volume; P-F, PaO2 to FiO2 ratio; PCO2, partial pressure of carbon dioxide; PLT, platelets; PO2, partial pressure of oxygen; PT, prothrombin time; PTT, partial thromboplastin time; RBC, red blood cell count; RDW CV, red cell distribution width coefficient of variation; RDW FL, red cell distribution width femtoliters; SBP, systolic blood pressure; TSH, thyroid stimulating hormone; WBC, white blood cell count; eGFR, estimated glomerular filtration rate.

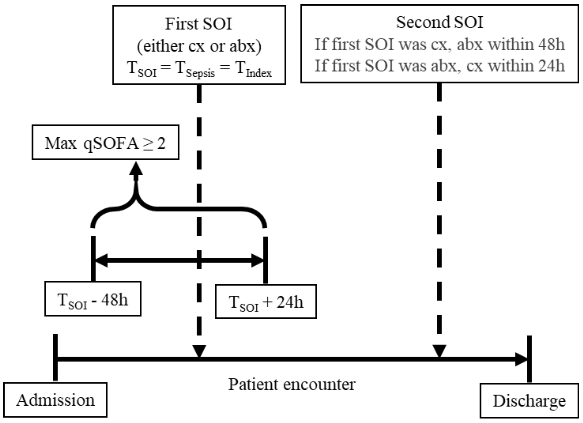
Appendix 35. Index time identification for sepsis and non-sepsis cohorts

For each patient encounter, a single index time was identified, and prediction was performed six hours prior to that index time.

For non-sepsis patients, index time was the maximum of [12 hours into admission] or [mid point between admission and discharge]:

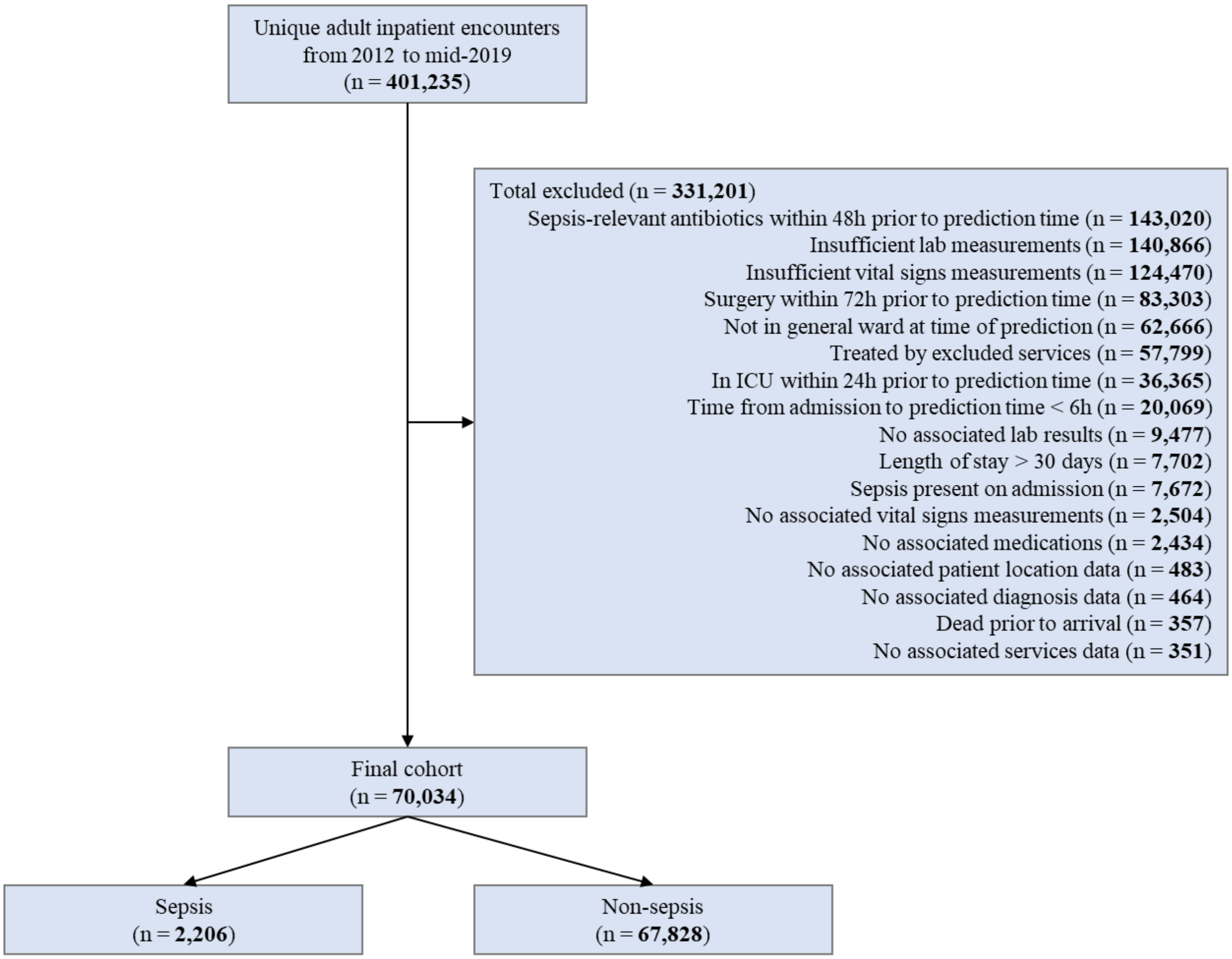


For sepsis patients, based on the elements that qualified the patient for meeting sepsis criteria, the time of suspicion of infection was also the time of sepsis and the time of index:



Abbreviations: SOI, suspicion of infection; qSOFA, quick sequential organ failure assessment; cx, cultures; abx, antibiotics.

Appendix 36. Cohort selection PRISMA-style diagram


Unique adult inpatient encounters from 2012 to mid-2019 (n = **401,235**)
Total excluded (n = **331,201**)
Sepsis-relevant antibiotics within 48h prior to prediction time (n = **143,020**)
Insufficient lab measurements (n = **140,866**)
Insufficient vital signs measurements (n = **124,470**)
Surgery within 72h prior to prediction time (n = **83,303**)
Not in general ward at time of prediction (n = **62,666**)
Treated by excluded services (n = **57,799**)
In ICU within 24h prior to prediction time (n = **36,365**)
Time from admission to prediction time < 6h (n = **20,069**)
No associated lab results (n = **9,477**)
Length of stay > 30 days (n = **7,702**)
Sepsis present on admission (n = **7,672**)
No associated vital signs measurements (n = **2,504**)
No associated medications (n = **2,434**)
No associated patient location data (n = **483**)
No associated diagnosis data (n = **464**)
Dead prior to arrival (n = **357**)
No associated services data (n = **351**)
Final cohort (n = **70,034**)
Sepsis (n = **2,206**)
Non-sepsis (n = **67,828**)

Each encounter can meet multiple exclusion criteria, thus the sum of number of encounters excluded by each criteria is greater than the total number of excluded encounters.

Appendix 37. Cohort characteristics: comorbidities

Variable	**Total** (n = 70,034)	**Sepsis** (n = 2,206)	**Non-sepsis** (n = 67,828)	**p**[a]
AIDS/HIV, n (%)	540 (0.8%)	11 (0.5%)	529 (0.8%)	0.173
Alcohol abuse, n (%)	1,120 (1.6%)	37 (1.7%)	1,083 (1.6%)	0.833
Blood loss anemia, n (%)	917 (1.3%)	31 (1.4%)	886 (1.3%)	0.759
Cardiac arrhythmias, n (%)	20,191 (28.8%)	1,142 (51.8%)	19,049 (28.1%)	< 0.01 *
Chronic pulmonary disease, n (%)	17,823 (25.4%)	782 (35.4%)	17,041 (25.1%)	< 0.01 *
Coagulopathy, n (%)	5,679 (8.1%)	505 (22.9%)	5,174 (7.6%)	< 0.01 *
Congestive heart failure, n (%)	19,846 (28.3%)	908 (41.2%)	18,938 (27.9%)	< 0.01 *
Deficiency anemia, n (%)	3,655 (5.2%)	132 (6.0%)	3,523 (5.2%)	0.111
Depression, n (%)	11,881 (17.0%)	419 (19.0%)	11,462 (16.9%)	0.011
Diabetes, complicated, n (%)	8,717 (12.4%)	377 (17.1%)	8,340 (12.3%)	< 0.01 *
Diabetes, uncomplicated, n (%)	13,099 (18.7%)	327 (14.8%)	12,772 (18.8%)	< 0.01 *
Drug abuse, n (%)	4,434 (6.3%)	97 (4.4%)	4,337 (6.4%)	< 0.01 *
Fluid and electrolyte disorders, n (%)	20,911 (29.9%)	1,300 (58.9%)	19,611 (28.9%)	< 0.01 *
Hypertension, n (%)	31,368 (44.8%)	1,067 (48.4%)	30,301 (44.7%)	< 0.01 *
Hypothyroidism, n (%)	9,001 (12.9%)	371 (16.8%)	8,630 (12.7%)	< 0.01 *
Liver disease, n (%)	6,061 (8.7%)	307 (13.9%)	5,754 (8.5%)	< 0.01 *
Lymphoma, n (%)	2,960 (4.2%)	124 (5.6%)	2,836 (4.2%)	< 0.01 *
Metastatic cancer, n (%)	6,676 (9.5%)	261 (11.8%)	6,415 (9.5%)	< 0.01 *
Obesity, n (%)	5,309 (7.6%)	245 (11.1%)	5,064 (7.5%)	< 0.01 *
Other neurological disorders, n (%)	4,622 (6.6%)	205 (9.3%)	4,417 (6.5%)	< 0.01 *
Paralysis, n (%)	1,861 (2.7%)	105 (4.8%)	1,756 (2.6%)	< 0.01 *
Peptic ulcer disease excluding bleeding, n (%)	566 (0.8%)	25 (1.1%)	541 (0.8%)	0.107
Peripheral vascular disorders, n (%)	5,514 (7.9%)	272 (12.3%)	5,242 (7.7%)	< 0.01 *
Psychoses, n (%)	1,243 (1.8%)	45 (2.0%)	1,198 (1.8%)	0.381
Pulmonary circulation disorders, n (%)	3,301 (4.7%)	241 (10.9%)	3,060 (4.5%)	< 0.01 *
Renal failure, n (%)	17,758 (25.4%)	769 (34.9%)	16,989 (25.0%)	< 0.01 *
Rheumatoid arthritis/collagen vascular diseases, n (%)	2,911 (4.2%)	111 (5.0%)	2,800 (4.1%)	0.042
Solid tumor without metastasis, n (%)	9,671 (13.8%)	384 (17.4%)	9,287 (13.7%)	< 0.01 *
Valvular disease, n (%)	7,699 (11.0%)	487 (22.1%)	7,212 (10.6%)	< 0.01 *
Weight loss, n (%)	5,987 (8.5%)	450 (20.4%)	5,537 (8.2%)	< 0.01 *

Appendix 38. XGBoost hyperparameter optimization

Random search on the training set was used to optimize XGBoost hyperparameters.

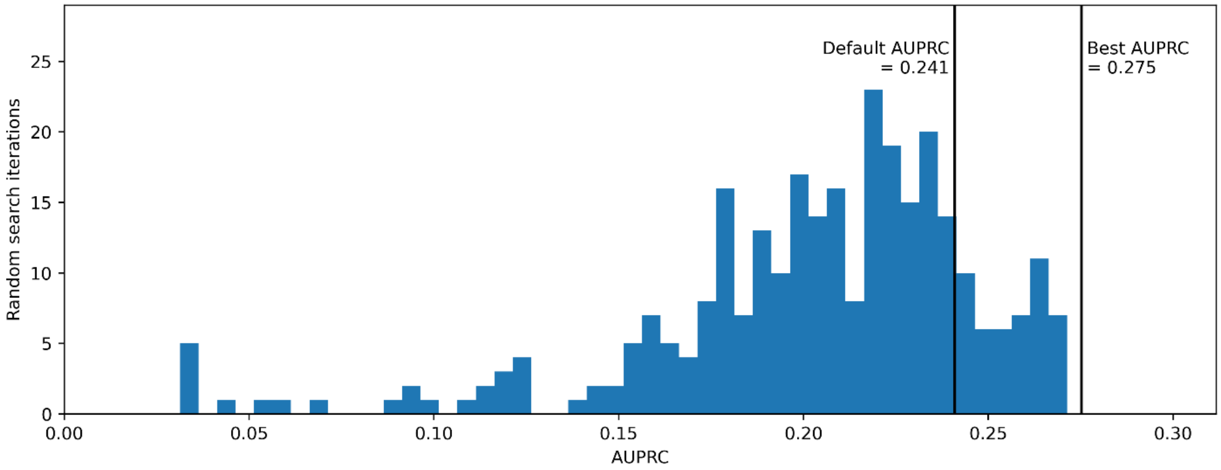
The fixed parameters were as follows:

Parameter	Value
tree_method	hist
grow_policy	depthwise
single_precision_histogram	True
n_estimators	100

The parameter spaces for optimization were as follows:

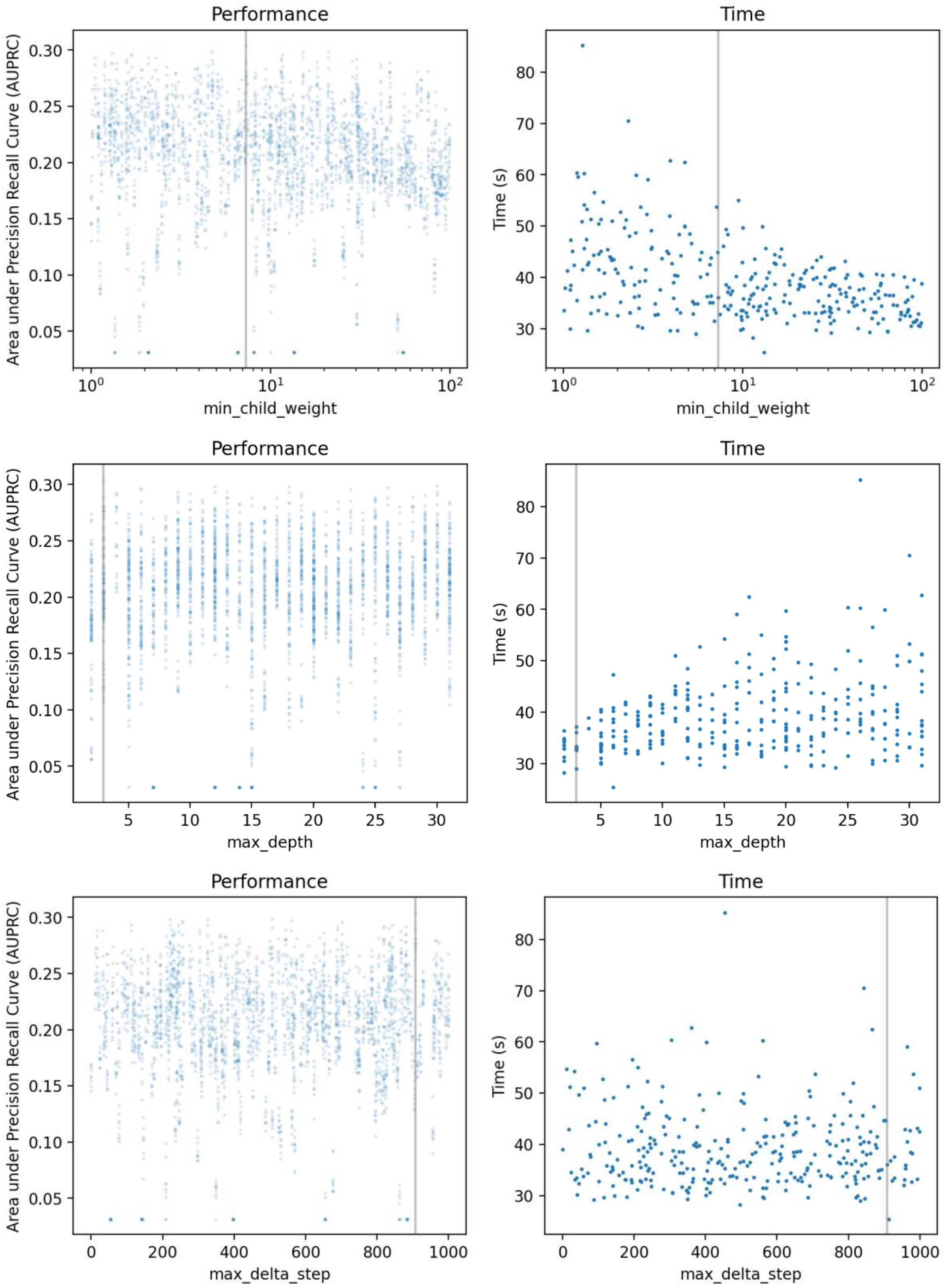
Parameter	Min	Max	Distribution
subsample	0.2	1.0	Uniform
colsample_bytree	0.2	1.0	Uniform
max_depth	2	32	Uniform
eta	1e-4	1	Log uniform
gamma	1e-2	1e2	Log uniform
max_bin	4	128	Uniform
min_child_weight	1	100	Log uniform
max_delta_step	0	1000	Uniform

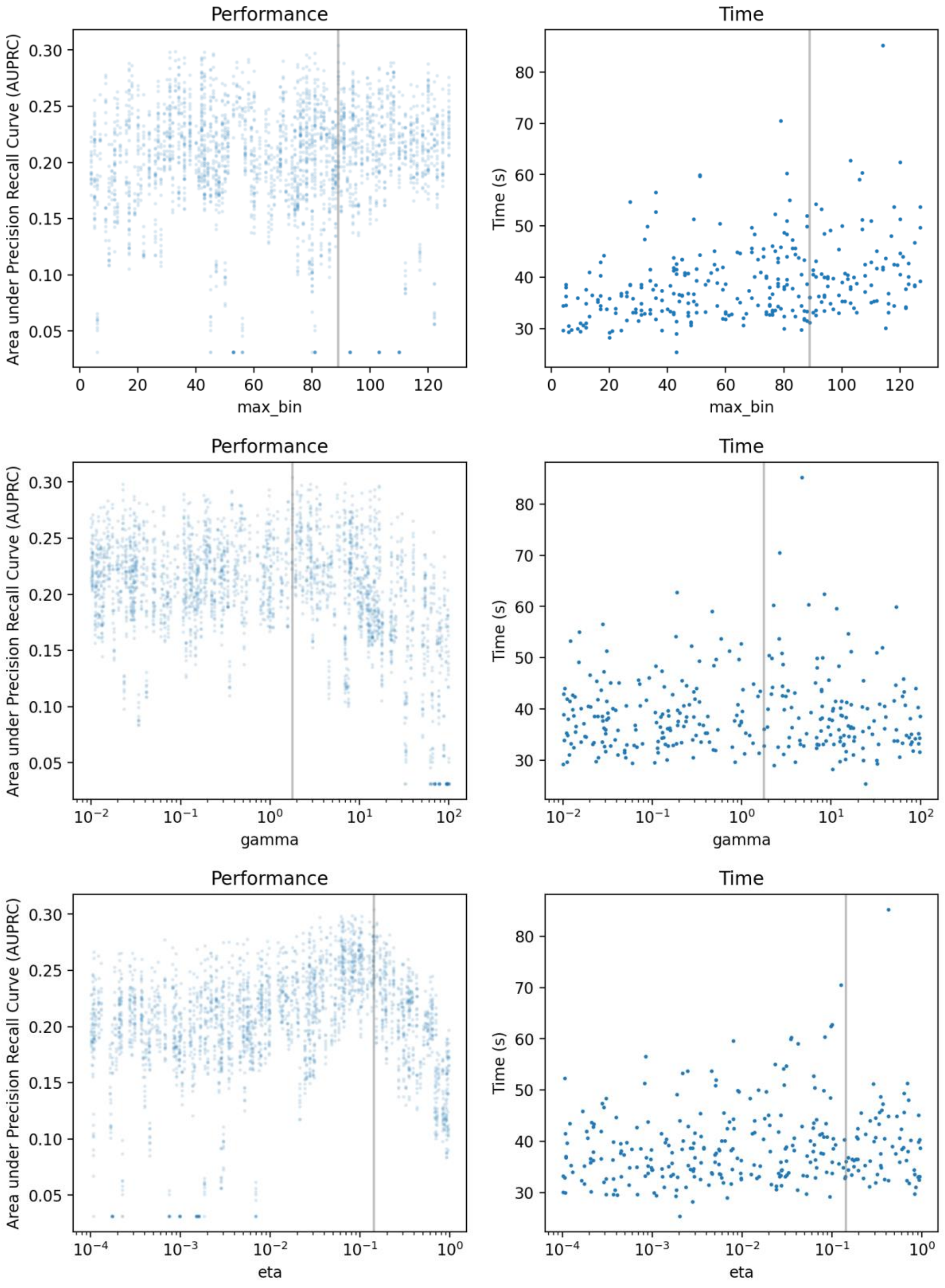
At each iteration, for each combination of parameters randomly sampled from the above distribution, 3-fold cross validation was repeated 3 times. Area under precision recall curve (AUPRC) was computed for each of the 9 splits. 300 iterations were performed yielding a total of 2,700 splits. The distribution of mean AUPRCs were as follows:

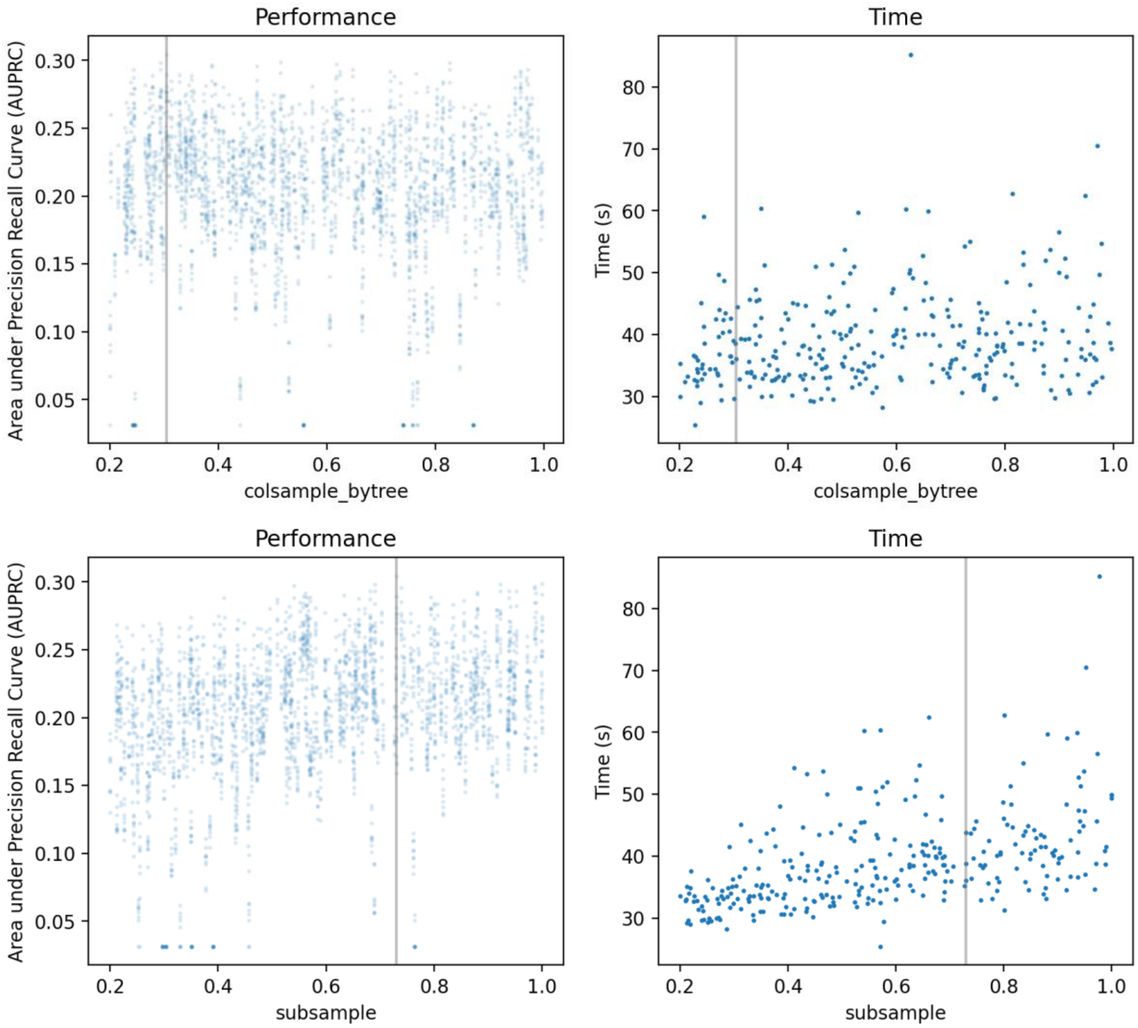


The best mean AUPRC found through random search was 0.275 compared to the median of 0.211 and 0.241 of an unoptimized XGBoost model using default parameters.

For each parameter, the parameter value was plotted against AUPRC and training time. For the AUPRC plots, each point represents a split whereas for the training time plot, each point represents the mean per iteration.


Performance
Time
Area under Precision Recall Curve (AUPRC)
0.30
0.25
0.20
0.15
0.10
0.05
Time (s)
80
70
60
50
40
30
10^0
10^1
10^2
min_child_weight
Performance
Time
Area under Precision Recall Curve (AUPRC)
Time (s)
5
10
15
20
25
30
max_depth
Performance
Time
Area under Precision Recall Curve (AUPRC)
Time (s)
0
200
400
600
800
1000
max_delta_step


Performance
Time
Area under Precision Recall Curve (AUPRC)
0.30
0.25
0.20
0.15
0.10
0.05
Time (s)
80
70
60
50
40
30
0
20
40
60
80
100
120
max_bin
Performance
Time
Area under Precision Recall Curve (AUPRC)
Time (s)
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
gamma
Performance
Time
Area under Precision Recall Curve (AUPRC)
Time (s)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
eta


Performance
Area under Precision Recall Curve (AUPRC)
0.30
0.25
0.20
0.15
0.10
0.05
0.2
0.4
0.6
0.8
1.0
colsample_bytree
Time
Time (s)
80
70
60
50
40
30
0.2
0.4
0.6
0.8
1.0
colsample_bytree
Performance
Area under Precision Recall Curve (AUPRC)
0.30
0.25
0.20
0.15
0.10
0.05
0.2
0.4
0.6
0.8
1.0
subsample
Time
Time (s)
80
70
60
50
40
30
0.2
0.4
0.6
0.8
1.0
subsample

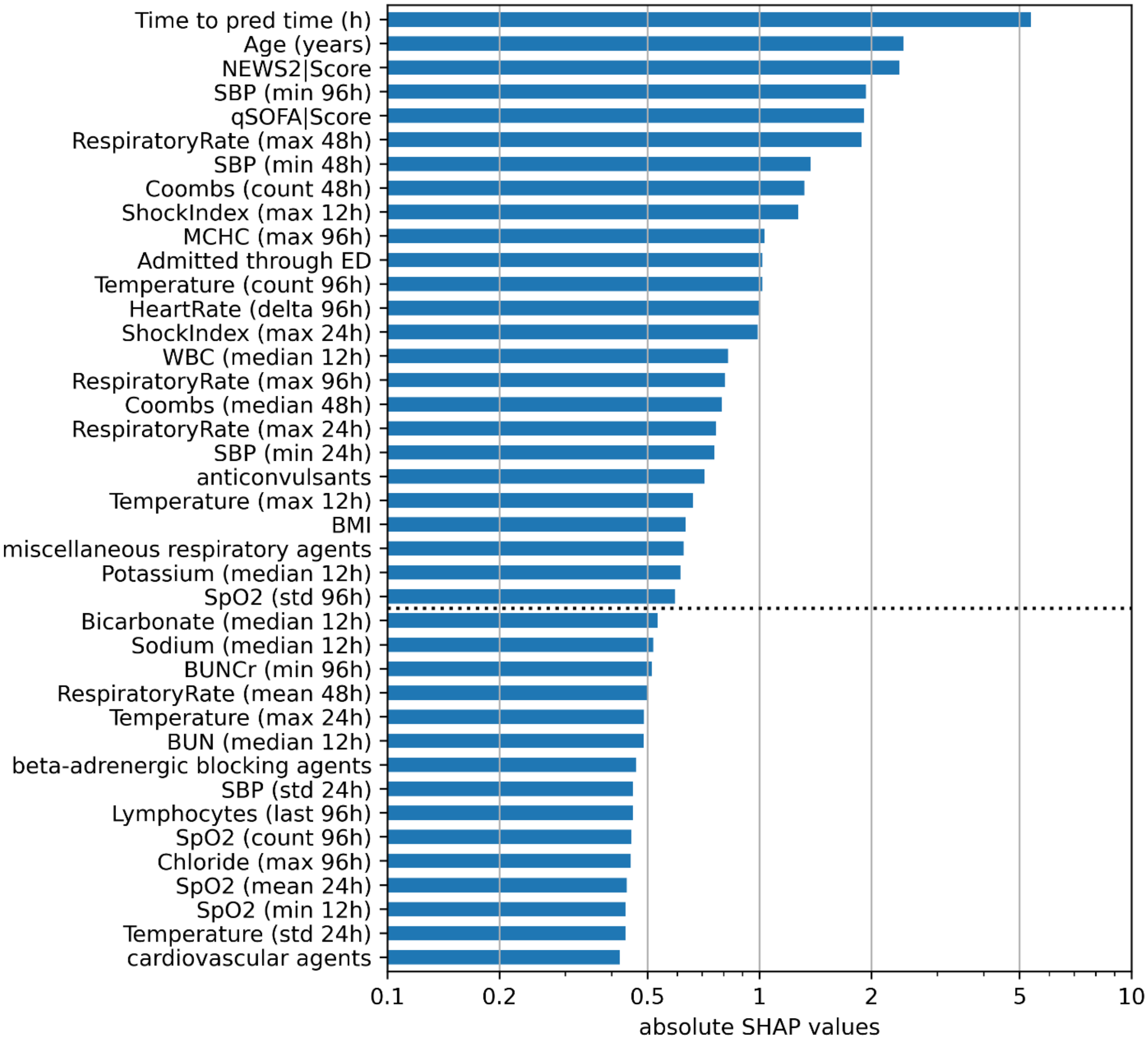
Appendix 39. SHAP values of optimized XGBoost model

Features	Absolute SHAP value (median [IQR])
Time to pred time (h)	0.262 (0.256 - 0.277)
Age (years)	0.125 (0.112 - 0.141)
NEWS2|Score	0.119 (0.097 - 0.132)
qSOFA|Score	0.098 (0.066 - 0.122)
SBP (min 96h)	0.097 (0.064 - 0.125)
RespiratoryRate (max 48h)	0.086 (0.071 - 0.115)
ShockIndex (max 12h)	0.063 (0.050 - 0.069)
Coombs (count 48h)	0.063 (0.047 - 0.089)
SBP (min 48h)	0.058 (0.047 - 0.080)
ShockIndex (max 24h)	0.052 (0.034 - 0.067)
Admitted through ED	0.051 (0.030 - 0.066)
MCHC (max 96h)	0.049 (0.030 - 0.072)
Temperature (count 96h)	0.047 (0.035 - 0.065)
HeartRate (delta 96h)	0.047 (0.040 - 0.059)
WBC (median 12h)	0.042 (0.035 - 0.046)
RespiratoryRate (max 96h)	0.039 (0.022 - 0.054)
Anticonvulsants	0.035 (0.022 - 0.048)
RespiratoryRate (max 24h)	0.035 (0.025 - 0.055)
Potassium (median 12h)	0.033 (0.021 - 0.042)
Temperature (max 12h)	0.033 (0.025 - 0.042)
Coombs (median 48h)	0.033 (0.027 - 0.054)
Miscellaneous respiratory agents	0.032 (0.022 - 0.041)
SBP (min 24h)	0.029 (0.015 - 0.057)
SpO2 (std 96h)	0.025 (0.013 - 0.045)
BMI	0.025 (0.014 - 0.045)

20 bootstrap samples were generated based on the training set, and for each feature, SHAP values (median and IQR) were calculated across bootstrap samples. The top 25 features based on absolute SHAP values are shown.

Abbreviations: NEWS2, National Early Warning Score 2; qSOFA, quick Sequential Organ Failure Assessment; SBP, systolic blood pressure; MCHC, mean corpuscular hemoglobin concentration; WBC, white blood cell; SpO2, oxygen saturation; BMI, body mass index.

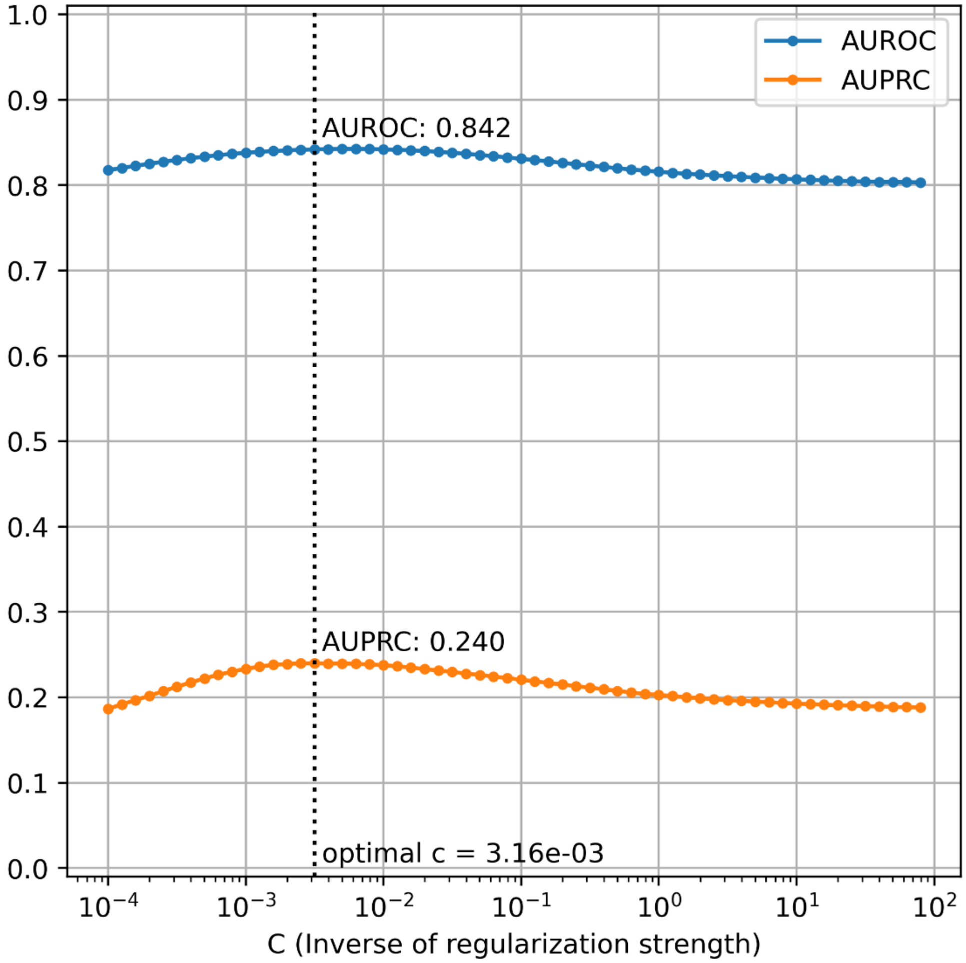
Appendix 40. Feature selection for lite model



20 bootstrap samples were generated based on the training set, and for each feature, absolute SHAP values were summed across bootstrap samples.

Based on relative drop-off in SHAP value, the cutoff (denoted by the dotted horizontal line) was drawn, and all of the features above the cutoff were used for the “lite” version of the XGBoost model (**XGB lite**).

Appendix 41. Logistic regression hyperparameter optimization



Using grid search with 3-fold, 3-repeat stratified cross validation on the training set, the optimal C (inverse regularization strength) parameter was searched between 10e-4 to 10e2 and was determined to be 3.16e-3, yielding a mean AUPRC of 0.240. This value of C was then used for the logistic regression model (**LogReg**).

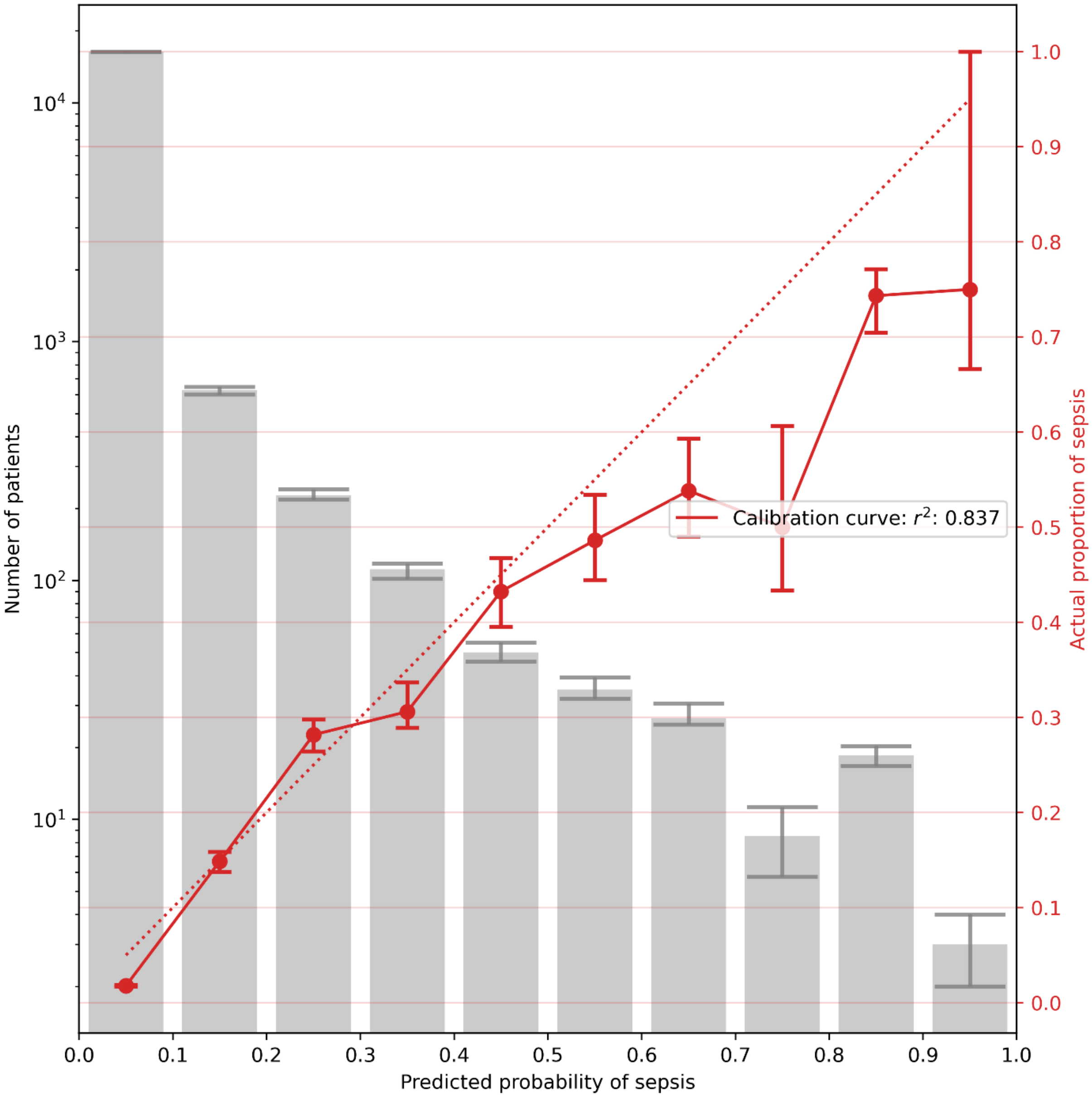
Appendix 42. Model performance comparison

Model	AUROC	AUPRC
XGB opt	**0.862 ± 0.011**	**0.294 ± 0.021**
XGB lite	0.856 ± 0.006	0.244 ± 0.013
XGB unopt	0.857 ± 0.007	0.287 ± 0.017
LogReg	0.857 ± 0.008	0.256 ± 0.024
NEWS2	0.699 ± 0.012	0.092 ± 0.009
qSOFA	0.705 ± 0.013	0.079 ± 0.006
SIRS	0.679 ± 0.010	0.066 ± 0.004

Model performance distributions were determined through 20 bootstrap samples on the test dataset.

Abbreviations: AUROC, area under receiver operating characteristic curve; AUPRC, area under precision recall curve; XGB opt, optimized XGBoost model; XGB lite, simple XGBoost model; XGB unopt, unoptimized, out-of-the-box XGBoost model; LogReg, logistic regression; NEWS2, National Early Warning Score 2; qSOFA, quick Sequential Organ Failure Assessment; SIRS, Systemic Inflammatory Response Syndrome.

Appendix 43. Calibration plot for optimized XGBoost model



For each of the 20 bootstrap samples on the test set, subjects were binned into deciles of predicted probability of sepsis. The grey bar plot and left y-axis represents the number of subjects in each bin (median and IQR). The red line plot and right y-axis represents the proportion of actual septic subjects in each bin (median and IQR).

Appendix 44. Pseudo-prospective trial, alert confusion matrix

	Sepsis	Non-sepsis	
Alerted	388	3144	PPV = 11.0%
Not alerted	169	13740	
	Sensitivity = 69.7%	Specificity = 81.4%	F1 = 19.0%

Based on the 17,441 encounters in the test dataset, after application of exclusions. Alerts for non-sepsis patients could be from any part of the patient encounter whereas alerts for sepsis patients can only be from before sepsis onset.

Appendix 45. Pseudo-prospective trial, time to intervention or outcome for alerted subjects

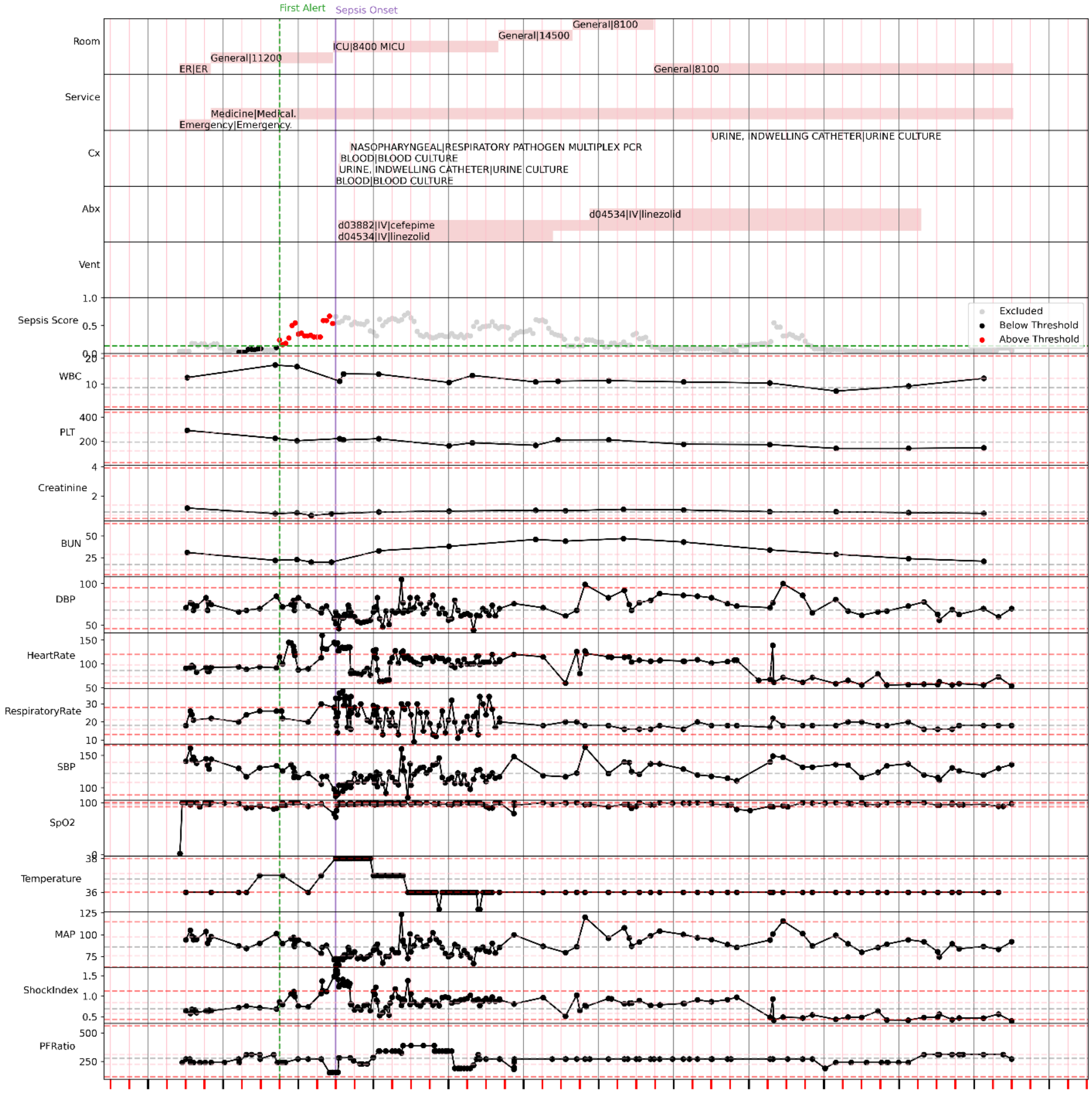
Intervention or Outcome	n (%)	Time to event (h), median (IQR)
Sepsis-relevant Cultures	1,376 (39.0%)	31.1 (11.4 - 75.2)
Sepsis-relevant Anti-infectives	991 (28.1%)	52.6 (20.8 - 115.7)
Ventilator Initiation	225 (6.4%)	65.5 (25.7 - 135.3)
Sepsis Onset	388 (11.0%)	29.8 (11.4 - 71.6)
ICU Transfer	371 (10.5%)	57.1 (17.8 - 128.5)
Death	164 (4.6%)	191.5 (81.5 - 320.7)

Appendix 46. Pseudoprospective trial, patient trajectory visualizations

Vertical solid blue line represents time of sepsis whereas the dotted blue line represents the first time in the encounter the predicted probability of sepsis crossed the threshold, which would have triggered an alert. Each black tick on the x-axis represents 24 hours whereas each red tick represents 6 hours.

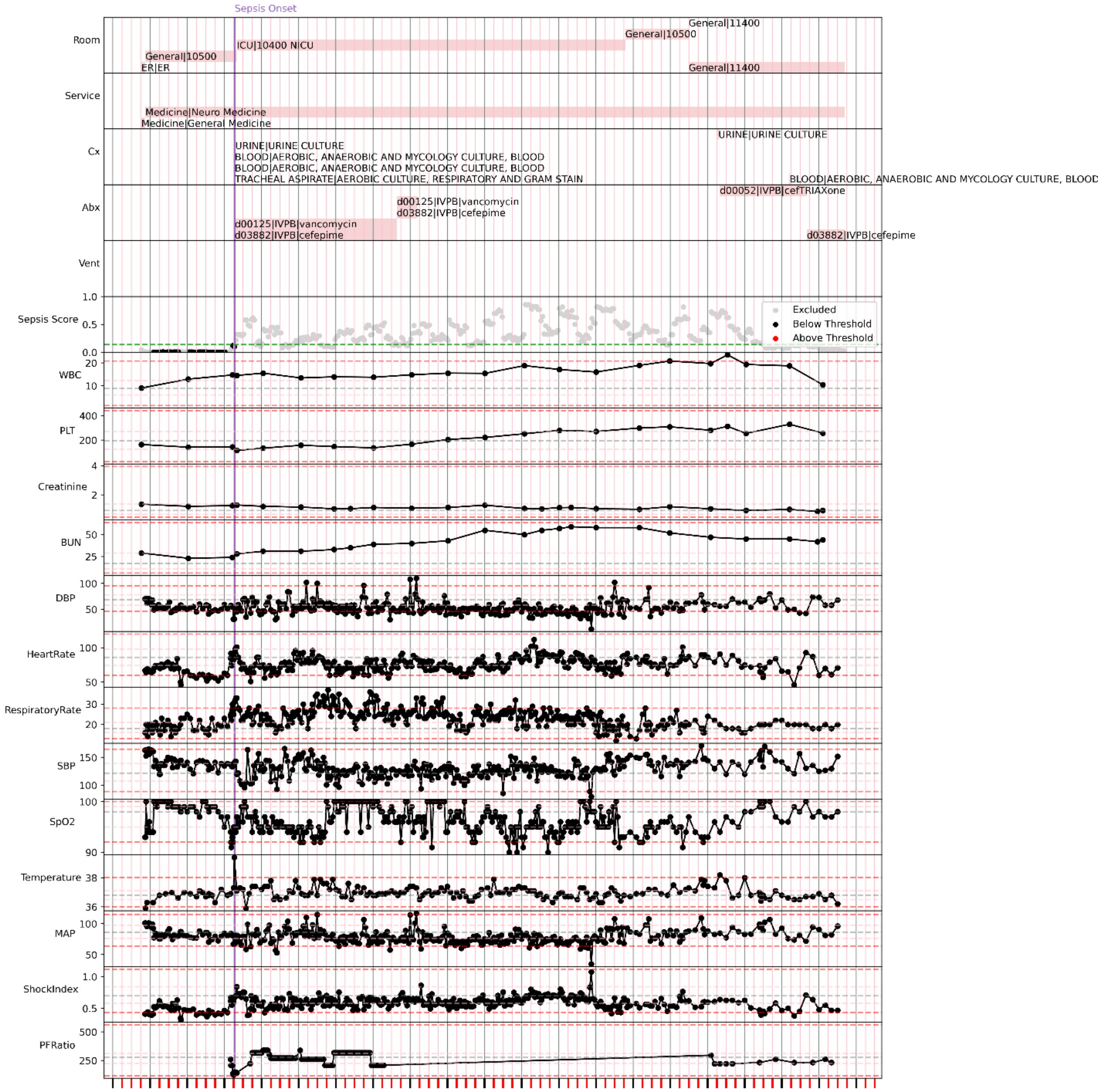
Abbreviations: Cx, sepsis-relevant cultures; Abx, sepsis-relevant anti-infectives; Vent, ventilator; WBC, white blood cell; PLT, platelets; BUN, blood urea nitrogen; DBP, diastolic blood pressure; SBP, systolic blood pressure; MAP, mean arterial pressure; PFRatio, PaO2 FiO2 ratio.

Example of alert success (true positive):



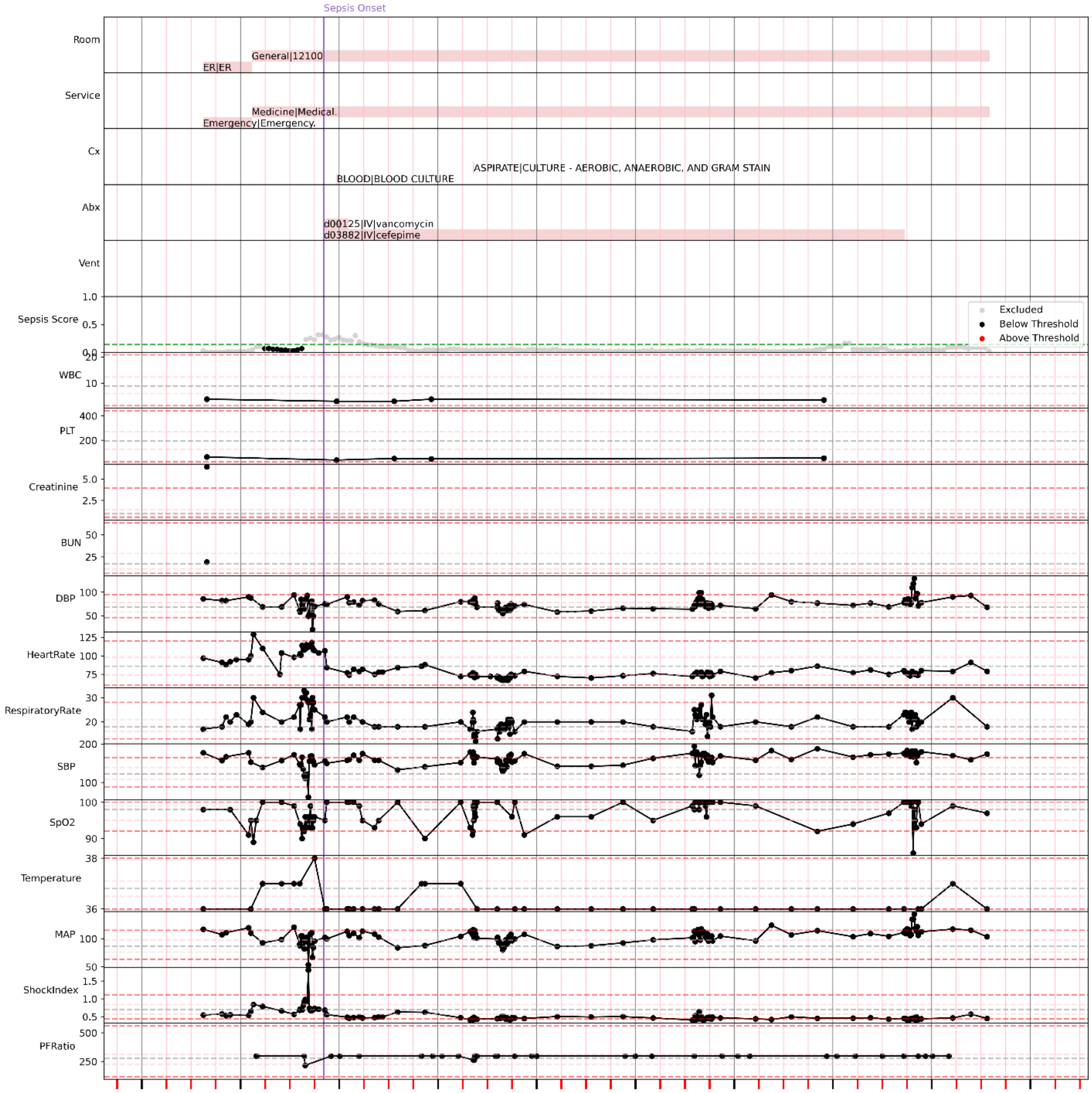
An alert fired roughly 18 hours prior to sepsis-relevant culture collection, sepsis-relevant anti-infective administration, sepsis onset, and ICU transfer.

Example of alert failure (false negative):



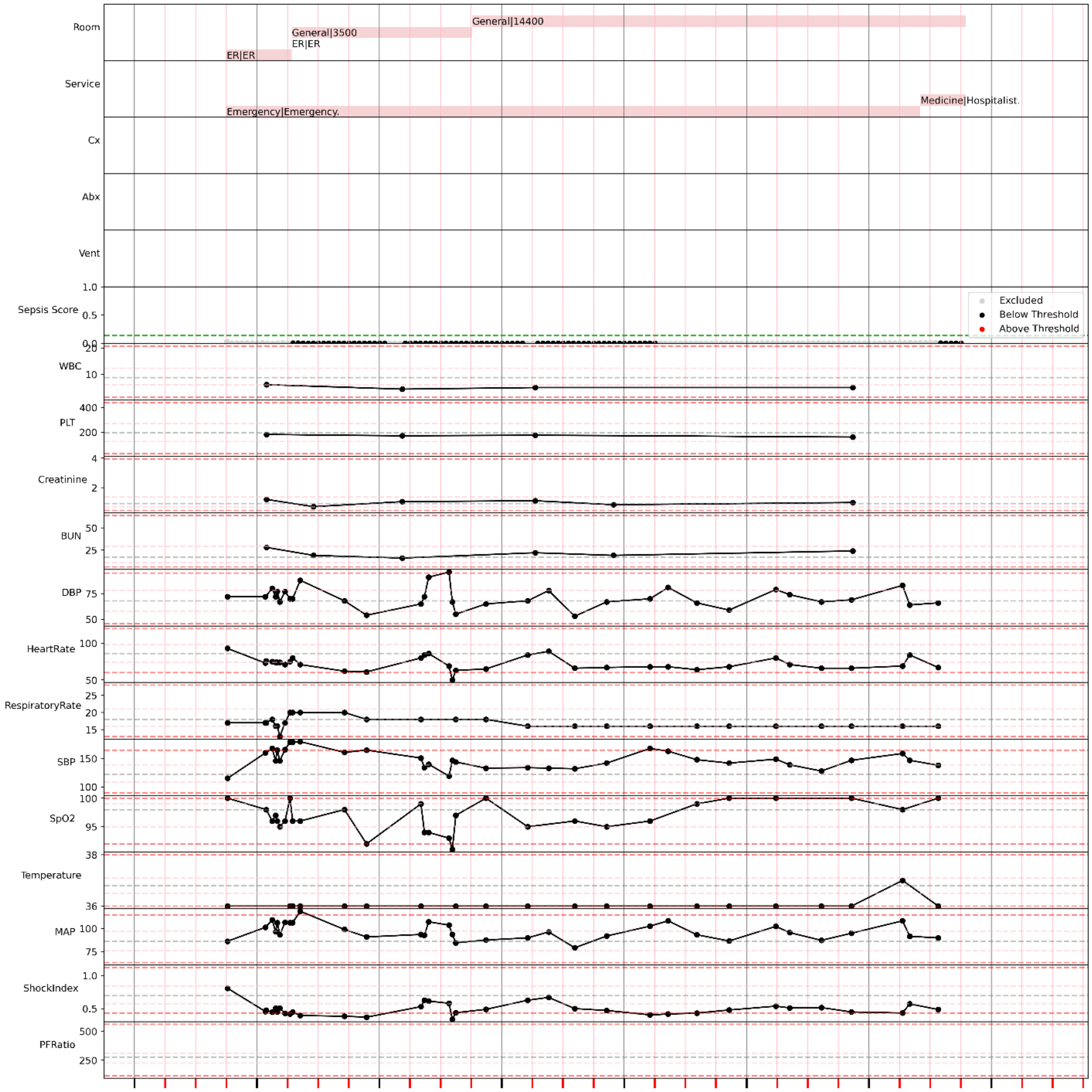
The sepsis risk score was consistently low up to the point of sepsis onset and ICU transfer.

Example of alert failure (false negative):



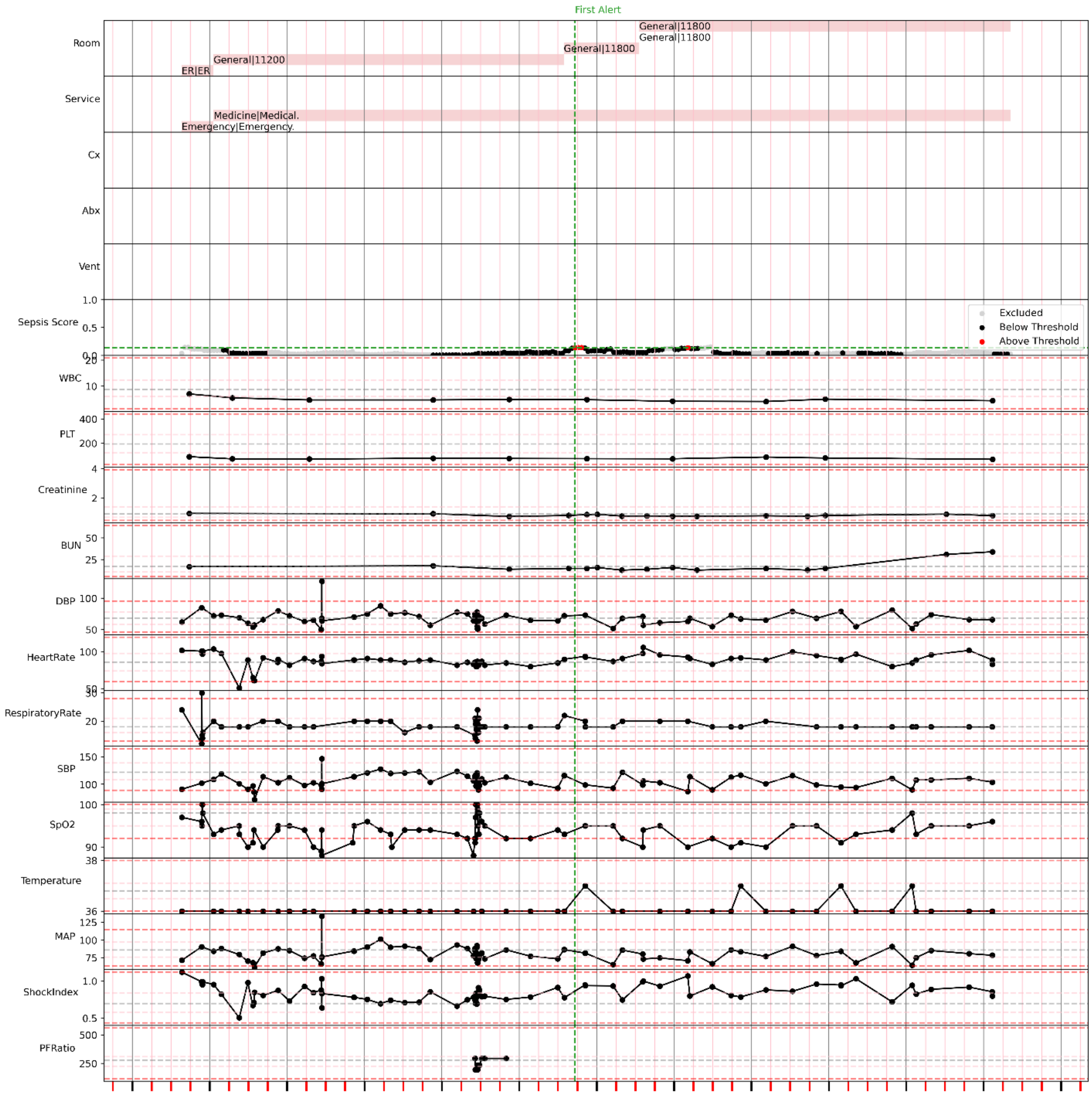
While there were scores crossing the threshold preceding sepsis onset, they were suppressed due to the lack of common labs (CBC/BMP) in the 24 hours preceding evaluation time.

Example of model success (true negative):



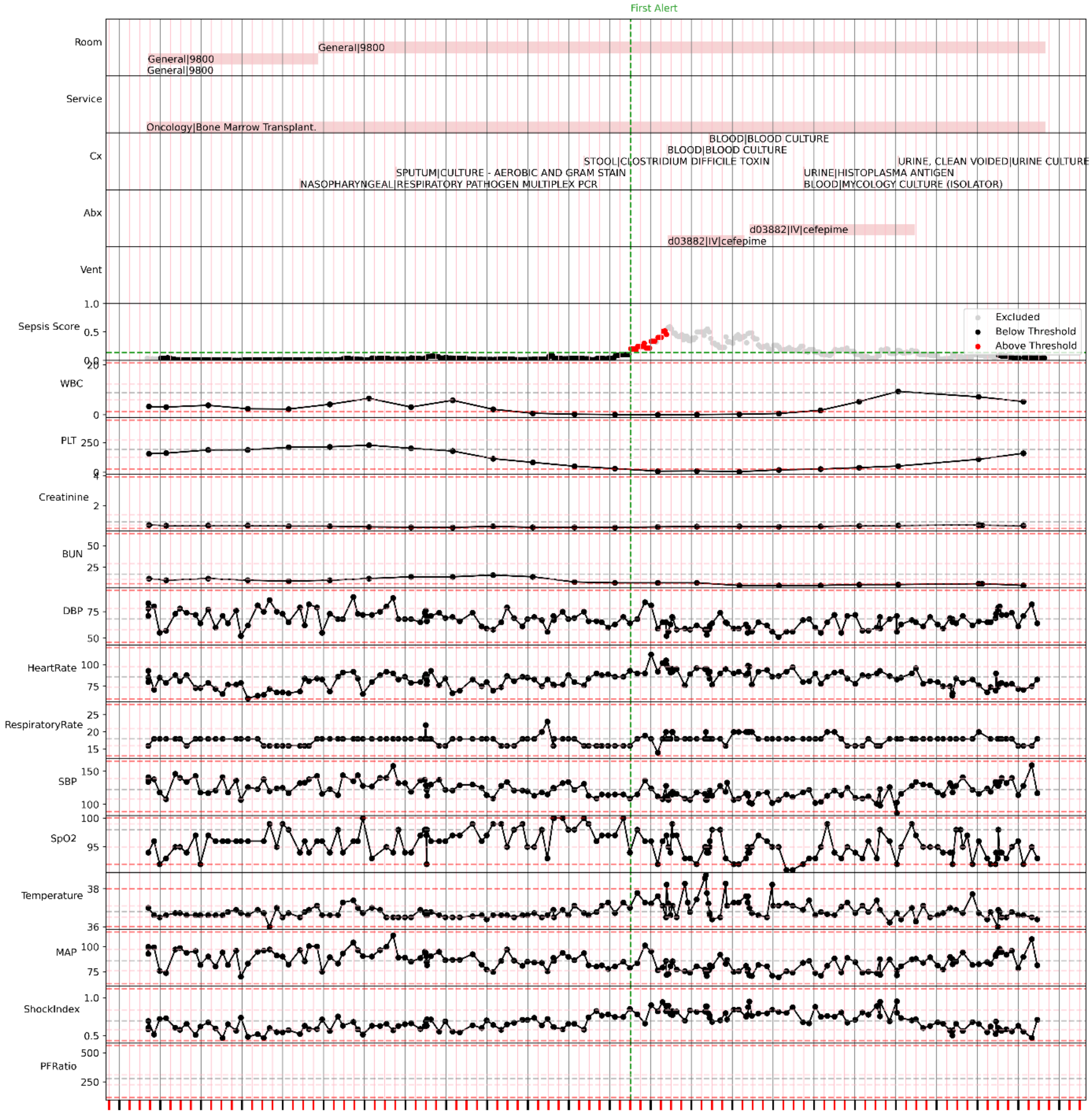
This patient was never septic, and the alert never crossed the threshold.

Example of alert failure (false positive):



While the patient was never septic, there were short periods where the score just barely exceeded the threshold.

Example of model failure (false positive):



While this patient was never septic (due to not meeting qSOFA), they did receive sepsis-relevant anti-infectives and cultures.